NeuroImage 129 (2016) 95–104

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Computational substrates of social norm enforcement by unaffected third parties

Songfa Zhong [a], Robin Chark [b], Ming Hsu [c,*], Soo Hong Chew [a,*]

[a] Department of Economics, National University of Singapore, Singapore 117570, Singapore
[b] Faculty of Business Administration, University of Macau, Avenida da Universidade, Taipa, Macau, China
[c] Haas School of Business and Helen Wills Neuroscience Center, University of California, Berkeley, CA 94720-1900, USA

ARTICLE INFO

Article history:
Received 16 March 2015
Accepted 18 January 2016
Available online 26 January 2016

Keywords:
Social norms
Altruistic punishment
fMRI
Ventromedial prefrontal cortex

ABSTRACT

Enforcement of social norms by impartial bystanders in the human species reveals a possibly unique capacity to sense and to enforce norms from a third party perspective. Such behavior, however, cannot be accounted by current computational models based on an egocentric notion of norms. Here, using a combination of model-based fMRI and third party punishment games, we show that brain regions previously implicated in egocentric norm enforcement critically extend to the important case of norm enforcement by unaffected third parties. Specifically, we found that responses in the ACC and insula cortex were positively associated with detection of distributional inequity, while those in the anterior DLPFC were associated with assessment of intentionality to the violator. Moreover, during sanction decisions, the subjective value of sanctions modulated activity in both vmPFC and rTPJ. These results shed light on the neurocomputational underpinnings of third party punishment and evolutionary origin of human norm enforcement.

Introduction

Social norms, the shared understandings of actions that are obligatory, permitted, or forbidden, play a central role in human societies in regulating social behavior, maintaining social coherence, and promoting cooperation (Bendor and Swistak, 2001; Camerer, 2003; Elster, 1989; Fehr and Fischbacher, 2004; Ostrom, 2000). In particular, the ability to develop norms and enforce them through the use of sanctions is thought by many to be one of the distinguishing characteristics of the human species (Boyd, 1988; Fehr and Fischbacher, 2003). The sanction may be either through reciprocal means taken by individuals whose economic payoff is directly harmed by the norm violation, or through impartial bystanders, so called "third parties", who are unaffected by the deviation but in a position to punish the violator (Bendor and Swistak, 2001; Fehr and Fischbacher, 2004; Ostrom, 2000).

In the case of reciprocal punishment, notable progress has been made in our understanding of its neural substrates through application of functional neuroimaging techniques to experimental games that capture core cognitive processes underlying norm-guided behavior (De Quervain et al., 2004; Knoch et al., 2006; Li et al., 2009). Using economic game paradigms such as the ultimatum game, these studies have identified critical roles for the insula cortex and anterior cingulate cortex (ACC), which are previously known to encode the emotion of disgust and conflict resolution respectively, in responding to norm violation in various settings (Sanfey et al., 2003; Xiang et al., 2013).

In addition, these studies have suggested that regions in the frontoparietal circuits to be important for assessment of intentionality and responsibility. Dorsolateral prefrontal cortex (DLPFC), for example, has been shown to be important in assessing intentionality of norm violation (Buckholtz et al., 2008; Haushofer and Fehr, 2008), and that their disruption via rTMS causally affects norm-related decisions (Buckholtz et al., 2015; Knoch et al., 2006). Studies of social behavior also reveal the right temporoparietal junction (rTPJ) in mentalizing and theory of mind, the ability to take perspectives from others (Frith and Frith, 2006). Finally, reward-related regions including striatum and ventromedial prefrontal cortex (vmPFC) have also been implicated social reward processing and sanctioning behavior (De Quervain et al., 2004; Knoch et al., 2006; Li et al., 2009).

In contrast, despite its ubiquity and importance to norm enforcement in human societies, we know much less in the case of enforcement by impartial bystanders (Bendor and Swistak, 2001; Fehr and Fischbacher, 2004; Ostrom, 2000). This has important implications for our understanding of the computational underpinnings of norm-guided behavior and their evolutionary origins (Fehr and Fischbacher, 2004; Riedl et al., 2012). Evolutionarily, humans constitute the only species known to have individuals regularly sanction norm violations even when they themselves are not affected, whereas reciprocal punishment is observed in multiple social species (Fehr and Fischbacher, 2004; Riedl et al., 2012). It has been suggested in the literature that both reciprocal punishment and third party punishment are crucial to the

* Corresponding authors.
E-mail addresses: mhsu@haas.berkeley.edu (M. Hsu), ecscsh@nus.edu.sg (S.H. Chew).

http://dx.doi.org/10.1016/j.neuroimage.2016.01.040
1053-8119/

establishment and maintenance of social norm (DeScioli and Kurzban, 2009, 2013). In addition, both types of punishment similarly depend on the extent of violation imposed on the offended as well as the intentionality of the violation on the part of the offender (Blount, 1995; Falk et al., 2003). That is, humans are capable of norm enforcement based on impartial community-based notions that are sensitive to the perspectives of the offender as well as the offended, which could be critical to both third party punishment and reciprocal punishment.

This is opposed to an alternative view that reciprocal punishment could be instead driven by non-norm-based concerns, such as retaliatory motives in response to status challenges, or simply "lashing out" (Fehr and Fischbacher, 2004; Riedl et al., 2012; Yamagishi et al., 2012). For example, under the "wounded pride hypothesis", reciprocal punishment such as rejection of unfair behavior in the ultimatum game results from a psychological response to a challenge to the integrity or inferior status of the responder (Yamagishi et al., 2012). By and large, current studies of reciprocal punishment are unable to differentiate between these explanations and have great difficulty accounting for sanctions by impartial bystanders (De Quervain et al., 2004; Sanfey et al., 2003; Xiang et al., 2013).

This, however, poses a challenge for current models of norm-guided behavior widely used in the studies of reciprocal punishment (Sanfey et al., 2003; De Quervain et al., 2004; Xiang et al., 2013). Specifically, norm-violations in these models are measured by so-called "egocentric inequity", defined as the difference between the absolute payoff difference between the decision-maker and other parties. That is, people are assumed to care about norm violation only to the extent their own relative position is affected. Note that the term "egocentric" refers only to the use of one's self as the frame of reference, as opposed to other colloquial meaning of selfishness. Thus, an important question for current neuroscientific accounts of social norms and norm-guided behavior is the extent to which computational components implicated in reciprocal punishment reflect the sophisticated capacities for norm enforcement by unaffected third parties (Montague and Lohrenz, 2007; Spitzer et al., 2007; Buckholtz et al., 2008). In addition, to what extent do computational demands involved in assessing norm violation from the perspective of others rely upon and recruit additional neural systems? And finally, how are norm-related computations from the perspectives of both offended and offending parties integrated to drive sanction behavior in unaffected third parties?

Here we adopt a set of third party punishment (TPP) games to probe the computational substrates of norm enforcement from the perspective of an impartial bystander. Specifically, we introduced a third party into the widely-used dictator game (DG) and scanned participants in the role of the third-party to investigate the neural responses to three key components of third party punishment: (1) how a third party responds to inequity between the dictator and the recipient, (2) how a third party responds to inequity when giving the option to punish the dictator, and (3) how a third party responds differently when the intentionality of the dictator differs. In this game, the dictator (P1) is given an endowment of 100 monetary units (MU), and can distribute any proportion of this endowment between herself and a recipient (P2). The dictator's decision is then revealed to the third party (P3). The third party, who is endowed with 160 MUs, must decide whether to sanction the behavior of the dictator at a ratio of 1:5. That is, for every MU spent by the third party, the dictator's earning is reduced by five MUs (Fig. 1A). Critically, to manipulate the perspective of the norm violator, we included, in addition to the standard TPP, a "No-Intention" condition where the distribution between the dictator and the recipient was decided by a randomization device rather than the dictator. That is, whereas in the standard "Intention" condition, any unfair distribution is the result of the dictator's choice, in the No-Intention condition, unfair distributions are the result of a random computer assignment. All other aspects of the game are identical between the conditions (Fig. 1A).

This paradigm has three important advantages as a cognitive probe of norm-guided behavior. First, unlike the ultimatum game and the trust game, the third party in this game does not stand to material gain or lose from the actions of the dictator. As a result, it is difficult for status or reciprocity motivated responses to account for observed sanctions. Most importantly, the parameters that the third party is endowed with more tokens than P1 were chosen such that standard egocentric models of norm enforcement would predict no punishment for all possible situations, including those that result in substantial inequity between the dictator and the recipient, thereby allowing us to separate egocentric and impartial motivations in observed sanction behavior. In addition, with a ratio of 1:3, it was observed that 40% of subjects choose no punishment for inequity distribution (Fehr and Fischbacher, 2004). As such, we use a higher ratio of 1:5 to better reveal heterogeneous preference for punishment. In addition, the temporal structure of the game enabled us to characterize not only the regions involved in processing key variables underlying behavior. More specifically, we are able to separately examine evaluation of the severity of norm violation when the P1's choice is first revealed to the third party in the Allocation event, and computation of subjective value of sanctioning said violations when the third party decides the level of punishment in the Sanction event.

Materials and methods

Subjects

22 right-handed student subjects (12 females, mean age 22.9 ± 3.2) were recruited through internet advertisements at Beijing Normal University. Of these subjects, one subject had excessive motion, and 3 subjects did not punish for all the trials. These four subjects were excluded from both behavioral and neuroimaging analyses.

Procedure

Subjects undergoing neuroimaging completed 24 rounds in one scanning session lasting 15–20 min. Each subjects' informed consent was obtained via consent form approved by the Internal Review Board at the Hong Kong University of Science and Technology and Beijing Normal University. Subjects in the scanner played the role of the third party, and were matched with 24 pairs of P1 and P2 who were selected from pretest experiments. Half the trials are under the Intention condition with the other half under No-Intention condition. The order of appearance of the two kinds of trials was randomized. The distributions of 100 MUs between P1 and P2 included 50:50, 80:20, 90:10 and 100:0 for both conditions. In particular, subjects were told that they were playing with real people for each round and that we would randomly match him/her with one pair of P1 and P2 only. Both P1 and P2 were paid after the fMRI experiment. The third party was informed that they would be paid based on one randomly chosen round from the 24 rounds plus a RMB160 participation fee. This method, widely used in fMRI experiment involving social interaction, adheres to the no-deception principle in experimental economics (De Quervain et al., 2004; Spitzer et al., 2007). This one-shot nature of the game ensures that there is no reputation effect, and it is incentive compactable for subjects to reveal their preference.

FMRI scanning parameters

The experiment was conducted by SIEMENS MAGNETOM Trio Tim 3 T MRI scanner. The echo spacing is 0.46 ms, EPI factor is 64, RF pulse type is normal, and gradient mode is fast. Subjects lay supine with their heads in the scanner bore and observed the rear-projected computer screen via a 45° mirror mounted above subjects' faces on the head coil. Subjects' choices were registered using two MRI-compatible button boxes. High-resolution T1-weighted scans ($1.3 \times 1.0 \times 1.3$ mm) were acquired on Siemens 3 T scanners. Functional images details: echo-planar imaging; repetition time (TR) = 2000 ms;

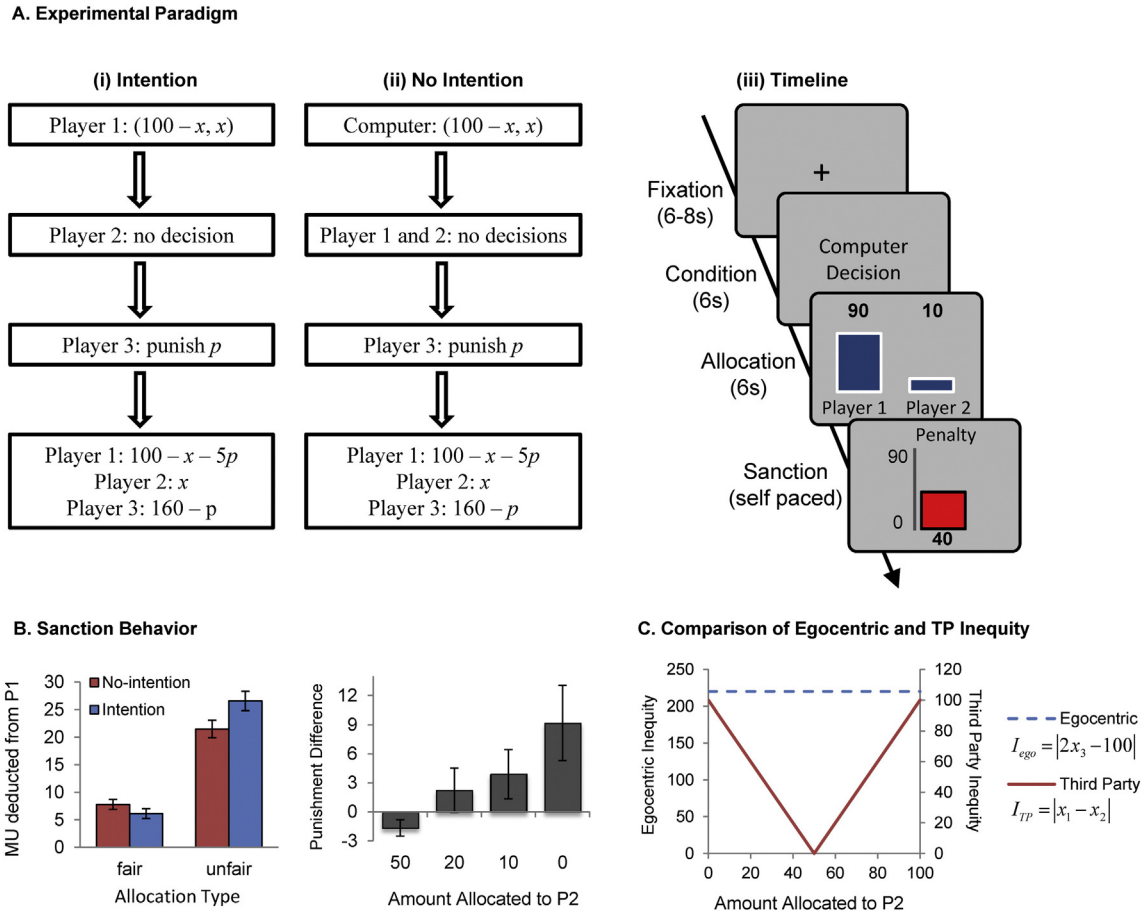

Fig. 1. (A) In the Intention condition, a dictator (P1) is given an endowment of 100 monetary units (MU), and can distribute any proportion of this endowment between P1 and a recipient (P2). The dictator's decision is then revealed to the third-party (P3), who decides whether to sanction the behavior of the dictator at a ratio of 1:5. That is, for every MU spent by the third-party, the dictator's earning is reduced by five MUs. The No-Intention condition is identical except the distribution between the dictator and the recipient is decided by a randomization device. The distributions of P1 and P2 are matched between two conditions (B) Third-party sanction decisions are modulated by both distribution inequity as well as intentionality. Left panel shows sanction varying between distributional inequity and intentionality. Fair trials are classified as 50:50 allocations, and unfair trials as trials where P1 receives 80 MUs or greater. Right panel shows paired differences between sanction in Intention and No-Intention conditions under different levels of distributional inequity. (C) Under egocentric models of inequity, the third party experiences the same level of inequity (blue dashed line) in the current game regardless of the amount that was allocated to P2 (x-axis). In contrast, under impartial, third party models of inequity, the third party experiences high levels of inequity (solid red line) when P2's allocation significantly departs from a 50–50 split of the initial endowment.

echo time (TE) = 30 ms; flip angle = 90° and functional $3.4 \times 3.4 \times 4$ mm^3 voxels.

FMRI data preprocessing

All the imaging data were processed and analyzed using SPM8 (Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, London — http://www.fil.ion.ucl.ac.uk/spm) and visualized in xjView (http://www.alivelearn.net/xjview8/). Functional images were realigned using a six-parameter rigid-body transformation. Each individual's structural T1 image was co-registered to the average of the motion-corrected images using 12-parameter affine transformation. Individual T1 structural images were segmented into grey matter, white matter, and cerebrospinal fluid before the individual grey matter was nonlinearly warped into MNI grey matter template. Functional images were, in order, slice-timing corrected, motion corrected, normalized into MNI space, and smoothed with an 8 mm isotropic Gaussian kernel.

FMRI data analysis

Random effects models were done in SPM8 by specifying a separate general linear model for each subject and pooled at the second level. All images were high-pass filtered in the temporal domain (filter width 128 s). Autocorrelation of the hemodynamic responses was modeled as an AR(1) process. Analyses of fMRI time series were done by generating distributional inequity, intentionality and decision utility from the computational model calibrated on choices of subjects at the individual level. An event-related design was used where regressors were included for the Allocation and Decision events of the trials (Fig. 1A). That is, for each subject, we constructed a (first level) general linear model (GLM) consisting of two events: an event at the time of Allocation with duration of 2 s, and one at the time of Decision with duration of 2 s. Regressors were convolved with the canonical hemodynamic response function and entered into a regression analysis against each subject's BOLD response data. The regression fits of each signal from each individual subject were then summed across their roles and then taken into random-effects group analysis.

For small-volume correction analysis, we use coordinates from previous studies within a 10-mm sphere (De Quervain et al., 2004; Greene et al., 2004; Mitchell, 2008; Sanfey et al., 2003). More specifically, we used coordinates of left insula (MNI coordinate, x = 35, y = 15, z = 3), right insula (MNI coordinate, x = −33, y = 14, z = −1) and ACC (MNI coordinate, x = 4, y = 20, z = 36) from the Sanfey et al. (2003), where activities in these regions were positively correlated with inequity between the proposer and the recipient in the ultimatum game. We adopted coordinates of left rDLPFC (MNI coordinate, x = −22, y = 48, z = 8) and right rDLPFC (MNI coordinate, x = 28, y = 49, z = 6) from an earlier study of the moral judgment (Greene et al., 2004), where rDLPFC was more activated when comparing the utilitarian with non-utilitarian moral judgment. We adopted the coordinates for

our SVC analysis from De Quervain et al. (2004)'s study of reciprocal altruistic punishment (MNI coordinate, x = 2, y = 54, z = −4). Finally, the rTPJ coordinate is based on a review of the theory of mind by Mitchell et al. (MNI coordinate, x = 54, y = −51, z = 27)(Mitchell, 2008). All these SVC results passed a corrected significance threshold of $p<0.05$.

Behavioral data analysis

Linear regression is used to test the effect of inequity on level of punishment, and standard errors are adjusted for clusters at the individual level. Structural estimation with logit specification is used to estimate the parameters of the model of third party punishment explained in next section. For trial t of subject i, for punishment level p_{it} given the set of feasible punishment levels ϑ,we specify the logit choice probability

$$P_{it}(x_1,x_2,x_3,p_{it},\overline{\gamma}) = \frac{\exp(u(x_1,x_2,x_3,p_{it},\overline{\gamma}))}{\sum_{\theta\in\vartheta}\exp(u(x_1,x_2,x_3,\theta,\overline{\gamma}))}.$$

The log-likelihood is specified as follows,

$$L(\overline{\gamma}) = \sum_{i=1}^{N}\sum_{t=1}^{T}\ln[P_{it}(x_1,x_2,x_3,p_{it},\overline{\gamma})].$$

We use maximum likelihood to estimate $\overline{\gamma}\in\{\gamma_I,\gamma_{NI}\}$, and test whether the two parameters are significantly greater than zero and whether they are significantly different for the two conditions.

Results

Third-party sanction behavior

First, we investigated how third-party sanction behavior varied as a function of both the distributional norm violation imposed upon the recipient, captured by inequity between the dictator and the recipient, and intentionality, captured by the intentionality of dictator's action. Multiple regression analysis showed that both factors had highly significant effects on sanctioning decisions (Fig. 1B). Specifically, the amount allocated by the dictator was significantly correlated with the level of sanction in both the No-Intention and the Intention conditions ($\beta_I = 0.52, p<0.001$; $\beta_{NI} = 0.34, p<0.001$, Fig. 1B). Importantly, sanctioning behavior was sensitive to the intentionality of the dictator, with significantly higher sanctions observed in the Intention condition than the No-Intention condition (paired t-test, $p<0.001$).

Egocentric models of inequity aversion cannot explain third-party punishment behavior

In the standard egocentric models of inequity averse behavior, decision-makers are assumed to be averse to payoff inequity between self and others only. In its linear version (Fehr and Schmidt, 1999), consider three players indexed by $i\in\{1,2,3\}$ for P1 the dictator, P2 the recipient and P3 the third party and let $x=\{x_1,x_2,x_3\}$ denote the vector of monetary payoffs. The utility of the third party under egocentric inequity takes the form:

$$u(x_1,x_2,x_3,p) = x_3 - \alpha\cdot\sum_{j=1,2}\max\{x_j-x_3,0\} - \beta\cdot\sum_{j=1,2}\max\{x_3-x_j,0\}$$

where x_3 captures third party's material payoffs, $\alpha\cdot\sum_{j=1,2}\max\{x_j-x_3,0\}$ captures inequity aversion where others have more monetary payoffs than self, and $\beta\cdot\sum_{j=1,2}\max\{x_3-x_j,0\}$ captures inequity aversion where others have less payoffs than self. The decision-maker is averse to both types of inequity when both α and β are positive.

In our game, the third party always has more monetary payoffs than P1 and P2, and hence punishment will increase the inequity between self and P1, leading to no punishment. Specifically, punishment will increase the advantageous inequity between third party and P1, $(x_3\text{-}p\text{-}(x_1\text{-}5p))$, and decrease the advantageous inequity between third party and P2, $[(x_3\text{-}p)\text{-}x_2]$. The aggregate effect on advantageous inequity would be $(x_3\text{-}(x_1+x_2)+3p)$, which is smallest when $p=0$. In addition, punishment also reduces her own monetary utility, $(x_3\text{-}p)$. Put it together, punishment increases the overall egocentric inequity and decreases monetary payoffs, and thus the third party would choose minimal level of sanction 0. Therefore the egocentric inequity aversion utility could not account for third party punishment under fairly standard assumptions of inequity aversion (Fig. 1C).

Computational modeling of third party punishment behavior

Going beyond an egocentric perspective, we extend the inequity utility model (Fehr and Schmidt, 1999) to incorporate the perspective of a third party. Specifically, we assume the third party dislikes the distributional inequity between P1 and P2. As punishment is costly, she needs to trade-off between own payoff and the level of distributional inequity between P1 and P2 without being directly involved in the inequity comparison. This can be contrasted with the aforementioned model of reciprocal punishment where the enforcer is at the center of the inequity calculations. We specify the choice model as follows:

$$u(x_1,x_2,x_3,p) = (x_3-p) - \overline{\gamma}\cdot|(x_1-5p)-x_2|,$$

where $(x_3\text{-}p)$ represents the post-sanction earnings of the third-party, and $|x_1\text{-}x_2\text{-}5\cdot p|$ the post-sanction distributional inequity between dictator and recipient. The aversion to inequity is captured by the parameter $\overline{\gamma}=\gamma_k$, with $k=I$ corresponding to the intention condition and $k=NI$ for the no-intention condition. The model captures three essential computation of third party punishment. First, the third party computes the inequity between P1 and P2 as $|(x_1\text{-}5p)\text{-}x_2|$. Second, the third party assigns different weights to inequity based on intentionality, which is captured by parameter $\overline{\gamma}$. Third, the third party computes the overall utility of punishment, and choose a level of punishment p to maximize utility $u(x_1,x_2,x_3,p)$.

This model extends previous inequity aversion models to the setting of third party punishment. Firstly, the model would predict a higher level of punishment if there is more inequity between P1 and P2. Secondly, we allow for the possibility that the degree of inequity aversion can depend on the nature of intentionality for the inequity between dictator and recipient. Intuitively, the third party dislikes inequity more under Intention condition than No-Intention condition, which we test in the subsequent estimation. Using this model, we were able to capture sensitivity to both inequity and intentionality as found in the regression analysis above. Specifically, we found that the third-party was significantly inequity averse in both the Intention ($\gamma_I = 0.18, p<0.001$) and the No-Intention conditions ($\gamma_{NI} = 0.13, p<0.001$). More importantly, and consistent with the regression results above, we found that the third-party was significantly more inequity averse in the Intention condition than the No-Intention condition (paired t-test, $p<0.005$, two-tailed). That is, P3 exerted greater sanctions on P1 when the distribution was more inequitable and when the offers were made by intentional acts of P1 (Fig. 1C).

In addition to the above "scaling" model, we considered an additional model where intentionality exerts additional weight in an additive manner, such that,

$$u(x_1,x_2,x_3,p) = (x_3-p) - \gamma\cdot|x_1-5p-x_2-\overline{\omega}|,$$

We next used a bootstrap procedure to compare these two models by comparing their AIC values. Our prediction is that the scaling model would outperform the additive model. In particular, the right

panel of Fig. 1B shows that the difference in punishment between intention and non-intentional is modulated by the size of the inequity. This is inconsistent with a shift in the constant but consistent with a difference in scaling. Indeed, we found that the scaling model performed significantly better than the additive model ($p<0.001$). Specifically, the AIC of the additive model exceeded that of the scaling model only once over 10,000 iterations.

Brain regions modulated by inequity from third-party perspective

Taken together, the model-based characterization of behavior provides a means to capture neural signatures of (i) inequity from a third party perspective measured by $|(x_1\text{-}5p)\text{-}x_2|$, (ii) the extent to which inequity was incurred intentionally, captured by parameter γ, and (iii) subjective value of sanctions motivated by inequity and intentionality, captured by $u(x_1,x_2,x_3,p)$. In our neuroimaging data analysis, we separately examine neural correlates of these three aspects of computation. First, we focused on the distributional inequity revealed upon Allocation event, measured by the difference in payoffs between the dictator and the recipient. If brain regions implicated in inequity processing during reciprocal punishment, in particular insula cortex and the ACC (Sanfey et al., 2003; Xiang et al., 2013), are sensitive to norms that apply to the community at large, we should observe a significant correlation between activity in these regions and inequity between the dictator and the recipient. Alternatively, such responses should be absent if the computational role of these regions is restricted to egocentric notions of inequity, as egocentric inequity is constant in our setup. Consistent with a general detecting norm-violation hypothesis, we found that activity in the ACC and bilateral insula cortex was significantly positively correlated with distributional inequity between the dictator and the recipient (Fig. 2A; Table 1). In addition to ACC and insula, we find that activation in the precuneus is significantly positively correlated distributional inequity. This is consistent with the role of precuneus in first-person perspective taking (Cavanna and Trimble, 2006).

Table 1
Neural response for increasing distributional inequity during Allocation event.

Regions	Cluster size (k)	Voxel-level statistics		MNI coor.		
		T-val	p_{unc}	x	y	z
R insula	18	4.5	0	27	18	−3
L insula	5	4.39	0	−30	21	6
	253	5.04	0	−9	21	42
Anterior cingulate cortex		5.04	0	33	0	45
		5.01	0	12	15	51
L superior frontal gyrus	12	3.88	0.001	−21	54	0
R inferior frontal		3.83	0.001	−27	51	−6
Gyrus	78	4.84	0	45	9	36
L superior frontal gyrus	21	5.01	0	−21	15	69
	1523	7.16	0	−21	−66	39
L precuneus		6.89	0	−33	−54	48
		6.83	0	−6	−84	9
	263	6.23	0	−42	9	54
L middle frontal gyrus		5.61	0	−48	12	33
		4.69	0	−33	6	48
	647	6.21	0	30	−57	48
R inferior parietal lobule		5.41	0	30	−69	45
		5.41	0	21	−72	57
	64	5.16	0	24	−18	3
R thalamus		4.95	0	15	−21	12
		4.22	0	21	−9	9
R declive	10	4.34	0	39	−72	−27

In addition, we investigated neural responses to the inverse of inequity in our game, fairness, during the Allocation Event. We found that activity in the ventromedial prefrontal cortex (vmPFC) was associated with distributional equity (Fig. 2B; Table 2). This is in line with a number of studies demonstrating vmPFC in reward computations including

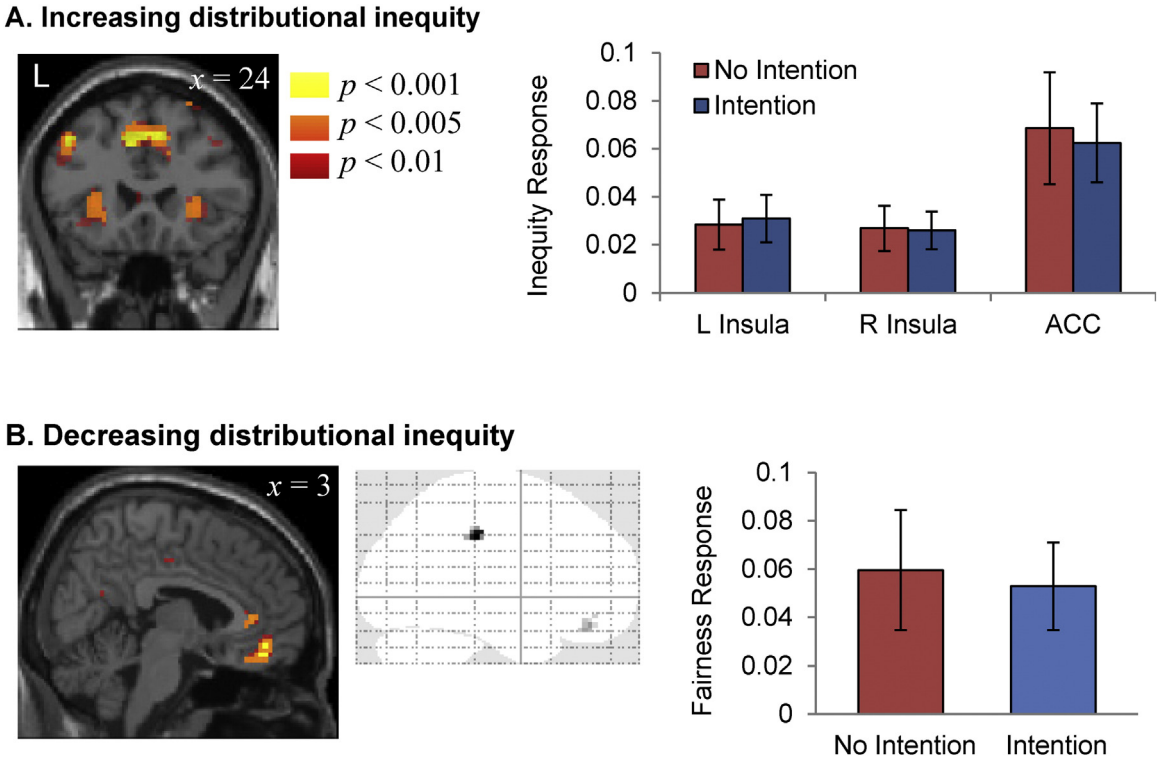

Fig. 2. (A) Bilateral insula and ACC where activities were significantly positively correlated with level of inequity pooling intention and no-intention condition (Left insula, MNI coordinate, x = −30, y = 21, z = 6; Right insula, MNI coordinate, x = 27, y = 18, z = −3; and ACC MNI coordinate, x = −9, y = 21, z = 42: $p<0.05$, small-volume-corrected, cluster size $k\geq 10$). Activities in these regions were not significantly modulated by intentionality ($p>0.5$). (B) Glass brain and sagittal section of vmPFC where activity is significantly negatively correlated with level of inequity pooling intention and no-intention condition (vmPFC, MNI coordinate, x = 3, y = 42, z = −18: $p<0.05$, small-volume-corrected, cluster size $k\geq 10$). Activity in this region was not significantly modulated by intentionality ($p>0.5$).

Table 2
Neural response for decreasing distributional inequity during Allocation event.

Regions	Cluster size (k)	Voxel-level statistics		MNI coor.		
		T-val	p_{unc}	x	y	z
Ventromedial PFC	18	3.96	0.001	3	42	−18
Cingulate gyrus	18	4.86	0	−9	−27	39

Table 3
Neural response for Intention > No-Intention during Allocation event.

Regions	Cluster size (k)	Voxel-level statistics		MNI coor.		
		T-val	p_{unc}	x	y	z
L anterior DLPFC	39	4.71	0	−27	41	10
		4.31	0	−21	53	13
R anterior DLPFC	21	4.31	0	21	56	13
R superior frontal gyrus	51	4.97	0	9	20	67
L occipital lobe	14	4.97	0	−9	−88	13
L cerebellum	12	4.13	0	−27	−55	−38

social ones (Hare et al., 2008; Li et al., 2009; O'Doherty et al., 2004; Padoa-Schioppa, 2007). Importantly, unlike in reciprocal punishment, the third-party stands neither direct monetary gain from equitable allocations, nor lose from inequitable allocations. Thus neural responses to equity (inequity) are more clearly reflective of impartial equity concerns *per se*, as opposed to either material gain (loss) or egocentric equity (inequity).

Brain regions selectively modulated by intentionality of norm violation

Next, we separated responses in the above regions of ACC and insula according to the Intention and No-Intention conditions, in order to investigate whether these neural inequity signals were also modulated by intentionality of the norm violation during Allocation event, the other key consideration underlying sanctioning decisions. If so, it would suggest that norm-related computations in these regions are also sensitive to the intentions of the norm violator. However, we found that inequity responses in the both ACC and insula were not significantly modulated by intentionality ($p>0.5$ for each).

In contrast, we found that bilateral anterior dorsolateral prefrontal cortex (DLPFC) responded to intentionality of the norm violation. Specifically, we contrasted neural responses between the Intention and No-Intention conditions during the Allocation event to localize regions with greater response in the Intention condition as compared to the No-Intention condition (Fig. 3A; Table 3). Moreover, this region appeared to be selective for intentionality. That is, using whole brain search, we found that activity in DLPFC did not respond to distributional inequity ($p>0.5$). Computationally, this may reflect neural processing related to assessment of the intentionality to the norm violator, and is consistent with previous neuroimaging results implicating this region in forming and updating beliefs about higher-order state associations (Burke et al., 2010; Gläscher et al., 2010). Indeed, we did not observe any significant activations in the reverse *No Intention*>Intention contrast even under liberal thresholds ($p>0.01$, Fig. 3B).

Brain regions modulated by subjective value of sanctions

Finally, we investigated how the brain integrates both inequity and intentionality signals in arriving at sanction decisions. The third party chooses an optimal level of punishment to maximize utility $u(x_1,x_2,x_3,p)$. For a chosenp^*, we can compute the subjective value of punishment measured by utility $u(x_1,x_2,x_3,p^*)$ for each trial, and examine the brain regions correlating with the subjective value of punishment. During the Sanction event, we found that none of the regions that responded to inequity, in particular ACC and insula cortex, was found to respond to subjective value of sanctioning decisions. Likewise, we did not observe a significant correlation with DLPFC activity. Instead, we found that responses in vmPFC and the right temporoparietal junction (rTPJ) were significantly correlated with the subjective value of sanctioning decisions ($p<0.05$, small-volume-corrected, $k\geq 10$, Fig. 4A; Table 4). Importantly, in follow-up region of interest analysis, we found that the vmPFC response was significantly greater in the Intention condition than in the No-Intention condition ($p<0.025$, Fig. 4B). In contrast, rTPJ response did not show a significant difference between the two conditions ($p>0.5$) (Fig. 4B).

Furthermore, we found that responses to the sanction value in vmPFC and rTPJ appeared specific to the Sanction event, as we did not observe significant vmPFC and rTPJ activity for sanction value during the Allocation event ($p>0.1$, uncorrected). Moreover, we did not find any brain region that responded to sanction value during the Allocation event even at liberal threshold ($p<0.01$, uncorrected). Together, these results suggest that computations relating to distributional inequity and intentionality are integrated only at the time of sanction decision.

Discussion

Norm enforcement by impartial third parties is thought to be crucial to the development of large-scale human societies, as norm compliance solely relying on a two-party retaliatory system typically cannot be sustained beyond the small-scale (Fehr and Fischbacher, 2004; Marlowe et al., 2008; Ostrom, 2000). In particular, enforcement through reciprocal punishment alone is known to fail when the cost of sanction by the violated is sufficiently high, or if the norm violation produces diffuse costs among many individuals. For example, in the case of cooperation norms, a shirking individual may impose little direct cost on any particular member, but result in collectively substantial damages. In such cases, norm enforcement by third parties can be crucial by sharing the cost of punishment beyond those directly affected by the norm violation (Bendor and Swistak, 2001; Fehr and Fischbacher, 2004). Consistent with this idea, behavioral experiments across small-scale societies have found that the level of punishment by third parties to be highly

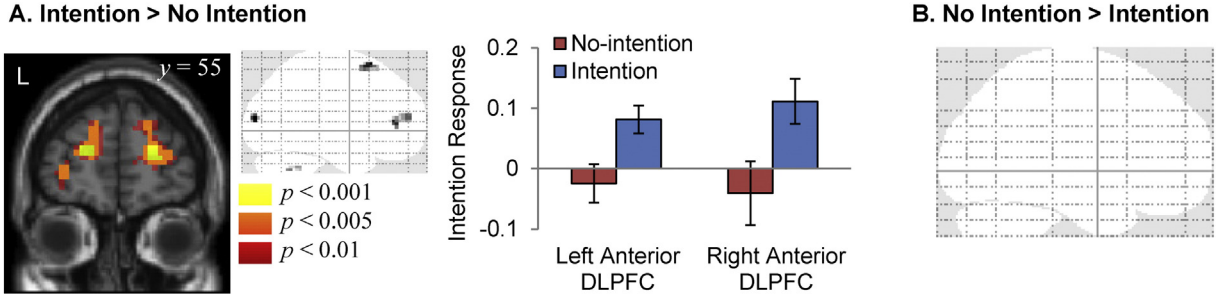

Fig. 3. (A) Glass brain and bilateral anterior DLPFC where activity is significantly greater in the Intention than No-Intention condition (left DLPFC, MNI coordinate, x = −27, y = 41, z = 10; right DLPFC, MNI coordinate, x = 21, y = 56, z = 13: $p<0.05$, small-volume-corrected, $k\geq 10$ voxels). (B) No brain region exhibited greater activation under No-Intention condition versus Intention condition ($p>0.01$, uncorrected).

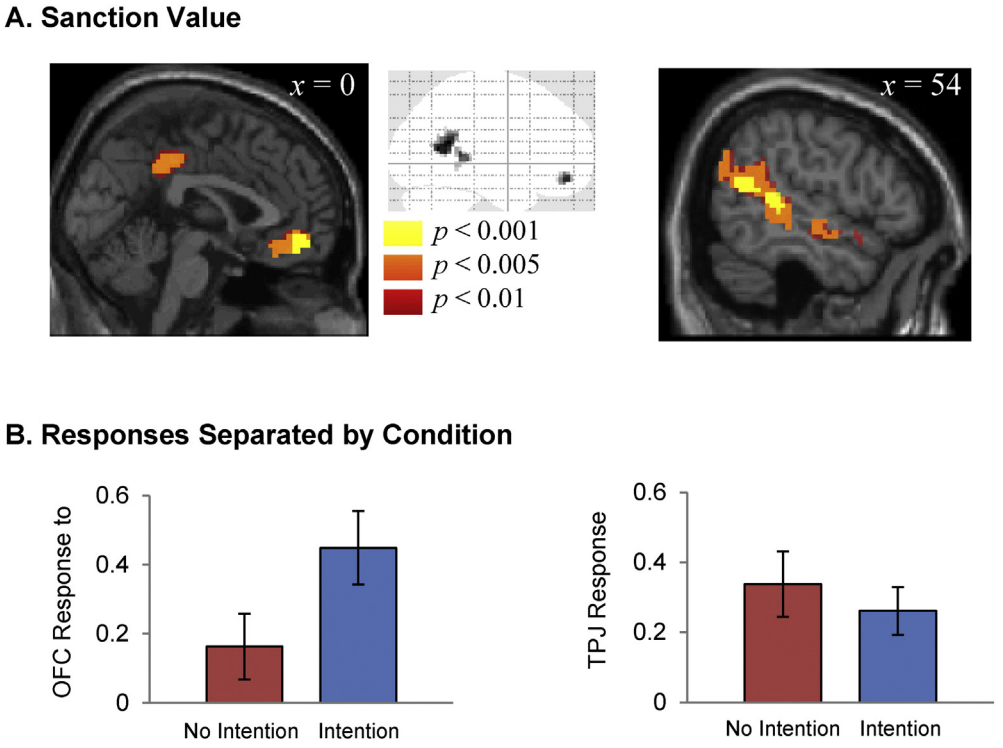

Fig. 4. (A) Glass brain and sagittal section of vmPFC and rTPJ where activity is significantly positively correlated with level of sanction value pooling intention and no-intention (vmPFC, MNI coordinate, x = 0, y = 47, z = −14; rTPJ, MNI coordinate, x = 54, y = −40, z = 7: $p<0.05$, small-volume-corrected, $k\geq10$). (B) Activity of vmPFC to sanction value is significantly greater in Intention condition than in No-Intention condition (paired t-test, $p<0.025$, two-tailed). Responsivity of rTPJ however is not significantly different for Intention condition and No-Intention condition (paired t-test,$p>0.5$, two-tailed).

correlated with the cooperation rates across societies (Henrich et al., 2006).

At the neural level, recent applications of functional neuroimaging methods, combined with computational models of inequity aversion, have transformed our understanding of the neural substrates of norm-guided behavior. Despite this progress, however, important questions remain about whether and to what extent these neural mechanisms reflect a general notion of norms that includes more general community concerns, or a narrower notion based on purely egocentric concerns. Here we address this question by characterizing the computational mechanisms underlying norm-guided behavior in third parties (Behrens et al., 2009; Hsu et al., 2008; Zhu et al., 2012). Specifically, using a set of third-party punishment games, we studied two factors driving third-party norm enforcement: (1) assessment of intentionality for and severity of norm violation, and (2) determination of appropriate level of sanction (Buckholtz et al., 2008; Fehr and Fischbacher, 2004; Spitzer et al., 2007).

First, despite clear psychological differences between egocentric and impartial notions of inequity, our results revealed a striking overlap between computational components involved in second and third party sanctions, and suggested that these regions exhibit the sophisticated capacities necessary for sensing norm violation in general. Specifically, we found opposing responses to inequity in ACC and insula on the one hand, and vmPFC on the other (Fig. 2). This is consistent with hypothesized functions of the former in processing aversive stimuli, including empathic pain and inequity (Chang and Sanfey, 2013; Civai et al., 2012; Corradi-Dell'Acqua et al., 2013; King-Casas et al., 2008; Sanfey et al., 2003; Singer et al., 2004), and the latter in reward processing (Hare et al., 2008; Li et al., 2009; O'Doherty et al., 2004; Padoa-Schioppa, 2007; Tabibnia et al., 2008; Tricomi et al., 2010; Zaki et al., 2013; Zaki and Mitchell, 2011). Critically, neural responses in these regions did not reflect egocentric notions of inequity or reward, but rather general notions of violation of distributional norm that are computed from the perspective of others. An important open question is the extent to which the current computational account could be generalized to cases where multiple norms coexist, such as when egocentric and impartial notions of inequity conflict. Indeed, in the present study, egocentric inequity was experimentally controlled to minimize conflicts between egocentric and impartial concerns. Future experiments can address this question by systematically manipulating relative payoff position and payoff distance among the players in a third party setting.

In contrast, we found that DLPFC but not ACC/insula differentiates intentionality in the Allocation event. This functional separation is consistent with previous evidence showing that assignment of responsibility is cognitively distinct from assessment of distributional norm violation (Buckholtz et al., 2008). One possible computational role for the DLPFC in such decisions is that it is involved in overriding automatic impulses to punish in the Intention condition (Haushofer and Fehr, 2008; Knoch et al., 2006). That is, because demands of impartiality may require overriding retributive motives in order to arrive at a reasonable judgment, the anterior DLPFC in our task may be involved in down-regulating psychological reward derived from punishment in the Intention condition. An alternative account, in contrast, posits that the DLPFC is involved in a constructive process that generates assessment of intentionality, for example, through abstract reasoning or belief maintenance (Fehr and Schmidt, 1999; Greene et al., 2004; MacDonald et al., 2000). This is consistent with findings in studies on social learning

Table 4
Neural response for decision utility during the sanction event.

Regions	Cluster size (k)	Voxel-level statistics		MNI coor.		
		T-val	P_{unc}	x	y	z
Ventromedial PFC	42	5.06	0	0	47	−14
R temporoparietal junction	41	4.88	0	54	−40	7
R superior temporal gyrus	142	5.33	0	60	−58	13
		5.48	0	63	−49	19

and model-based reinforcement learning, where DLPFC is implicated in computations related to forming and updating beliefs about higher-order state associations (Burke et al., 2010; Gläscher et al., 2010). Interestingly, both accounts are consistent with the finding that repetitive transcranial magnetic stimulation (rTMS) over the DLPFC reduces punishment of intentional norm violations. However, because only the latter posits a computational role for DLPFC in processing intent per se, it is possible to test these two accounts in future experiments where the intentionality of the dictator must be inferred or learned, rather than explicitly given as in the current case (Behrens et al., 2008; Hampton et al., 2008).

Notably, none of these regions, including ACC/insula and DLPFC identified in the Allocation event, appeared to respond to both distributional inequity and intentionality during the Sanction event. Instead, we found evidence of an integration and selection role for the vmPFC. During the Sanction event, the sanction value signal, which integrated distributional inequity and intentionality and the selected the level of punishment, was correlated with activity in the vmPFC. This region overlapped with the vmPFC activation identified in the Allocation event, but this activation cannot be accounted for by hemodynamic lag, as sanction value did not significantly correlate with vmPFC activity, nor respond to intentionality, during the Allocation event.

These results therefore support the view of norm enforcement as part of a hierarchical process whereby computations involving distributional inequity and of intentionality are integrated to arrive at sanctioning decisions in the vmPFC, and accord well with the hypothesized role of the vmPFC in neural representation, integration, and interaction of both monetary and social rewards (Behrens et al., 2009). The vmPFC is anatomically and functionally well suited to play this role, as it projects to several brain areas that are heavily involved in reward valuation, preference generation, and decision-making (Behrens et al., 2009; Hare et al., 2010; Hare et al., 2009; Hare et al., 2008; Rangel et al., 2008). Our findings regarding the vmPFC also echo those of previous studies in which investigators, using different paradigms, reported data suggesting that activations in a neural network including the vmPFC positively reinforce social rewards (Hare et al., 2008; Li et al., 2009; O'Doherty et al., 2004; Padoa-Schioppa, 2007; Tabibnia et al., 2008; Tricomi et al., 2010; Zaki et al., 2013; Zaki and Mitchell, 2011), and more generally social cognition (Gusnard et al., 2001; Miller and Cohen, 2001; Saxe, 2006).

Interestingly, we found a similar sanction value response in the rTPJ, although unlike vmPFC, this response did not differ significantly according to intentionality. Although widely implicated in studies involving mentalizing and perception of agency (Frith and Frith, 2006), rTPJ activation is not typically observed in studies involving social preferences. In studies of reward-guided behavior, activation of rTPJ is normally associated with computations of learning signals related to belief updating and higher-order state associations (Behrens et al., 2008; Hampton et al., 2008), and is thought to reflect two separate operations—interacting with an opponent whose internal states can be modelled (i.e., another human) and whose behavior is also relevant for guiding one's future actions (Gläscher et al., 2010). Therefore, one possible interpretation is that rTPJ is an additional computational system involved in third party punishment, as the computation of value of third-party needs shifting attention away from the self to focus on the needs of others (Frith and Frith, 2006; Lamm et al., 2007; Mitchell, 2008). That is, decisions to punish require the third party to focus on the desires and well-being of others rather than upon one's own economic payoffs (Haushofer and Fehr, 2008).

Notably, we found that precuneus is positively correlated with distributional inequity in the Allocation event, but not in the Sanction event. The precuneus is widely implicated in studies involving visuospatial imagery, episodic memory retrieval and self-processing operations (Cavanna and Trimble, 2006). In particular, in studies of self-processing tasks, precuneus is more activated when subjects read self-descriptive traits compared to non-self-descriptive traits (Kircher et al., 2000), as well as when subjects read stories written in the first-person in comparison with a third-person perspective (Vogeley et al., 2001). Therefore, one possible interpretation for our observation is that precuneus is involved in the perception of inequity aversion at the Allocation event from the perspective of one's self, while rTPJ is engaged in order to take the perspective of others to reach a punishment decision at the Sanction event. In a meta-analysis of decision making in the ultimatum game (Feng et al., 2015), precuneus is more activated when comparing unfair offers with fair offers. This suggests that precuneus is a commonly shared neural mechanism for self-other referencing for both reciprocal and third party punishment.

It is less clear, however, the specific nature of the mechanisms by which information regarding intentionality, norm violation, and sanctioning decision are integrated. The fact that our scaling model outperformed the additive model suggests DLPFC modulates, or "gates" an inequity aversion signal used to arrive at a punishment decision, as opposed to one where DLPFC exerting direct additional weight in punishment. However, we did not find strong evidence at the neural level that speak to the nature of DLPFC's involvement in punishment. Specifically, we tested the extent to which the individual intention parameter, γ_I-γ_{NI}, was correlated with differential neural activities between the Intention and No-Intention conditions in vmPFC, insula, DLPFC, and rTPJ. None of these tests were significant even at liberal thresholds ($p>0.1$ for all tests).

In light of these null findings, therefore, it may be desirable in future studies to causally manipulate DLPFC functioning instead of relying on the inherently correlational nature of fMRI measures. Indeed, there is some recent evidence using rTMS that are consistent with the gating hypothesis. In particular, application of rTMS to the DLPFC in legal judgment was found to reduce punishment by simultaneously diminishing the influence of information about culpability and enhancing the influence of information about harm severity (Buckholtz et al., 2015). Although seemingly paradoxical, this is consistent with our behavioral model, as well as the hypothesis that DLPFC encodes the information about culpability and gating the harm severity, which would lead to the reduced punishment when such responses are disrupted.

More generally, together with previous studies of reciprocal punishment and moral judgment, these results raise the intriguing possibility that involvement of theory of mind processes, subserved by frontoparietal circuits, may be a critical component accounting for the uniqueness of the human species in third party norm enforcement. That is, unlike reciprocal punishment, third-party norm enforcement requires individuals to represent norms from the perspective of others as opposed to one's self. Although necessarily speculative, this is consistent with recent nonhuman primate evidence, which found that chimpanzees, the closest living phylogenetic relative to humans, do not punish those who steal from third parties, even as they readily punish those who steal from them directly (Riedl et al., 2012). This is particularly relevant given causal evidence suggesting a necessary role of rTPJ in supporting a uniquely human cognitive capacity to represent and reason about mental states of others (Carter et al., 2012; Saxe, 2006; Young et al., 2010), which in turn may help to explain the unique nature of human engagement in third-party norm enforcement.

Limitations and conclusions

Finally, there are two important open questions concerning the external validity of our results specifically, and our conceptualization of norm violation more generally. The first concerns the widespread forms of sanctions involving non-pecuniary means. For example, social exclusion such as ostracism and corporal punishment are some of the most common forms of social sanctions in response to social norm violations (Guala, 2012). It would be of interest to examine whether and how individual difference in third party punishment game would predict actual behavior in non-pecuniary forms of punishment. Future studies combining altruistic punishment with manipulations of social

exclusion (Eisenberger et al., 2003) and physical punishment (McDermott et al., 2009) would be necessary.

The second concerns the psychological mechanism by which inequity aversion influences behavior. Whereas models of inequity aversion assume that decision-makers receive direct disutility from inequity, an alternative, non-mutually exclusive, approach is to allow players to have preferences regarding the beliefs of others, such that inequity affects behavior by shaping decision-makers' expectations (Battigalli and Dufwenberg, 2007; Dufwenberg and Kirchsteiger, 2004; Geanakoplos et al., 1989; Rabin, 1993). For example, Battigalli and Dufwenberg (2007) proposed that a model of "guilt aversion" where a player's social preferences depend on her beliefs about another player. Specifically, she feels "guilty" when she takes an action that deviates from others' expectations of her action, akin to "letting the other player down". In the third party punishment setting, it is possible that P3 punishes because he/she believes that is what P2 expects. Intuitively, this corresponds to a case where not punishing leads P3 to feel guilty about letting P2 down. Differentiating between this and our account, however, requires additional measures of beliefs, either using direct elicitation or manipulation of player beliefs. Future studies, such as combining our third-party punishment game and belief elicitation mechanisms used in guilt-aversion studies (Chang and Sanfey, 2013; Chang et al., 2011), would be necessary to address this important question.

Acknowledgments

We thank X.T. Zhu for assistance in data collection and the State Key Laboratory of Cognitive Neuroscience and Learning at Beijing Normal University for facilitating the fMRI experiment. This research was supported by the Hong Kong University of Science and Technology (S.Z. and S.H.C.) (AOE-MG/H-03/06), the Ministry of Education, Singapore (S.Z. and S.H.C.) (R122000230646; R122000237112), the National Institute of Mental Health (R01 MH098023, to M.H.), and the Hellman Family Faculty Fund (M.H.).

References

Battigalli, P., Dufwenberg, M., 2007. Guilt in games. Am. Econ. Rev. 97 (2), 170–176 (May 1).
Behrens, T.E., Hunt, L.T., Woolrich, M.W., Rushworth, M.F., 2008. Associative learning of social value. Nature 456, 245–249.
Behrens, T.E., Hunt, L.T., Rushworth, M.F., 2009. The computation of social behavior. Science 324, 1160–1164.
Bendor, J., Swistak, P., 2001. The evolution of norms. Am. J. Sociol. 106, 1493–1545.
Blount, S., 1995. When social outcomes arent fair: the effect of causal attributions on preferences. Organ. Behav. Hum. Decis. Process. 63, 131–144.
Boyd, R., 1988. Culture and the Evolutionary Process. University of Chicago Press.
Buckholtz, J., Asplund, C., Dux, P., Zald, D., Gore, J.C., Jones, O., Marois, R., 2008. The neural correlates of third-party punishment. Neuron 60, 930–940.
Buckholtz, J.W., Martin, J.W., Treadway, M.T., Jan, K., Zald, D.H., Jones, O., Marois, R., 2015. From Blame to Punishment: Disrupting Prefrontal Cortex Activity Reveals Norm Enforcement Mechanisms. Neuron 87 (6), 1369–1380 (Sep 23).
Burke, C.J., Tobler, P.N., Baddeley, M., Schultz, W., 2010. Neural mechanisms of observational learning. Proc. Natl. Acad. Sci. 107, 14431–14436.
Camerer, C., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press.
Carter, R.M., Bowling, D.L., Reeck, C., Huettel, S.A., 2012. A distinct role of the temporal–parietal junction in predicting socially guided decisions. Science 337, 109–111.
Cavanna, A.E., Trimble, M.R., 2006. The precuneus: a review of its functional anatomy and behavioural correlates. Brain 129 (3), 564–583 (Mar 1).
Chang, L.J., Smith, A., Dufwenberg, M., Sanfey, A.G., 2011. Triangulating the neural, psychological, and economic bases of guilt aversion. Neuron 70 (3), 560–572 (May 12).
Chang, L.J., Sanfey, A.G., 2013. Great expectations: neural computations underlying the use of social norms in decision-making. Soc. Cogn. Affect. Neurosci. 8, 277–284.
Civai, C., Crescentini, C., Rustichini, A., Rumiati, R.I., 2012. Equality versus self-interest in the brain: differential roles of anterior insula and medial prefrontal cortex. NeuroImage 62, 102–112.
Corradi-Dell'Acqua, C., Civai, C., Rumiati, R.I., Fink, G.R., 2013. Disentangling self-and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. Soc. Cogn. Affect. Neurosci. 8, 424–431.
De Quervain, D.J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., 2004. The Neural Basis of Altruistic Punishment. Science.
DeScioli, P., Kurzban, R., 2009. Mysteries of morality. Cognition 112, 281–299.
DeScioli, P., Kurzban, R., 2013. A solution to the mysteries of morality. Psychol. Bull. 139, 477.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Game Econ. Behav. 47 (2), 268–298 (May 31).
Eisenberger, N.I., Lieberman, M.D., Williams, K.D., 2003. Does rejection hurt? An fMRI study of social exclusion. Science 302 (5643), 290–292 (Oct 10).
Elster, J., 1989. Social norms and economic theory. J. Econ. Perspect. 3, 99–117.
Falk, A., Fehr, E., Fischbacher, U., 2003. On the nature of fair behavior. Econ. Inq. 41, 20–26.
Fehr, E., Fischbacher, U., 2003. The nature of human altruism. Nature 425, 785–791.
Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evol. Hum. Behav. 25, 63–87.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114, 817–868.
Feng, C., Luo, Y.J., Krueger, F., 2015. Neural signatures of fairness-related normative decision making in the ultimatum game: A coordinate-based meta-analysis. Hum. Brain Mapp. 36 (2), 591–602 (Feb 1).
Frith, C.D., Frith, U., 2006. The neural basis of mentalizing. Neuron 50, 531–534.
Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Game Econ. Behav. 1 (1), 60–79 (Mar 31).
Gläscher, J., Daw, N., Dayan, P., O'Doherty, J.P., 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66, 585–595.
Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D., 2004. The neural bases of cognitive conflict and control in moral judgment. Neuron 44, 389–400.
Guala, F., 2012. Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. Behav. Brain Sci. 35, 1–15.
Gusnard, D.A., Akbudak, E., Shulman, G.L., Raichle, M.E., 2001. Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. Proc. Natl. Acad. Sci. 98, 4259–4264.
Hampton, A.N., Bossaerts, P., O'Doherty, J.P., 2008. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc. Natl. Acad. Sci. 105, 6741–6746.
Hare, T.A., O'Doherty, J.P., Camerer, C.F., Schultz, W., Rangel, A., 2008. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J. Neurosci. 28, 5623–5630.
Hare, T.A., Camerer, C.F., Rangel, A., 2009. Self-control in decision-making involves modulation of the vmPFC valuation system. Science 324, 646–648.
Hare, T.A., Camerer, C.F., Knoepfle, D.T., Rangel, A., 2010. Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. J. Neurosci. 30, 583–590.
Haushofer, J., Fehr, E., 2008. You shouldn't have: your brain on others' crimes. Neuron 60, 738–740.
Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J., 2006. Costly Punishment Across Human Societies. Science (New York, N.Y.) 312, 1767–1770.
Hsu, M., Anen, C., Quartz, S.R., 2008. The right and the good: distributive justice and neural encoding of equity and efficiency. Science 320, 1092–1095.
King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., Montague, P.R., 2008. The rupture and repair of cooperation in borderline personality disorder. Science (New York, N.Y.) 321, 806–810.
Kircher, T.T., Senior, C., Phillips, M.L., Benson, P.J., Bullmore, E.T., Brammer, M., Simmons, A., Williams, S.C., Bartels, M., David, A.S., 2000. Towards a functional neuroanatomy of self processing: effects of faces and words. Cogn. Brain Res. 10 (1), 133–144 (Sep 30).
Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., Fehr, E., 2006. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314, 829–832.
Lamm, C., Batson, C.D., Decety, J., 2007. The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. J. Cogn. Neurosci. 19, 42–58.
Li, J., Xiao, E., Houser, D., Montague, P.R., 2009. Neural responses to sanction threats in two-party economic exchange. Proc. Natl. Acad. Sci. U. S. A. 106, 16835–16840.
MacDonald, A.W., Cohen, J.D., Stenger, V.A., Carter, C.S., 2000. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science 288, 1835–1838.
Marlowe, F.W., Berbesque, J.C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J.C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., 2008. More 'altruistic' punishment in larger societies. Proc. R. Soc. B Biol. Sci. 275, 587–592.
McDermott, R., Tingley, D., Cowden, J., Frazzetto, G., Johnson, D.D., 2009. Monoamine oxidase A gene (MAOA) predicts behavioral aggression following provocation. Proc. Natl. Acad. Sci. 106 (7), 2118–2123 (Feb 17).
Miller, E.K., Cohen, J.D., 2001. An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. 24, 167–202.
Mitchell, J.P., 2008. Activity in right temporo-parietal junction is not selective for theory-of-mind. Cereb. Cortex 18, 262–271.
Montague, P.R., Lohrenz, T., 2007. To detect and correct: norm violations and their enforcement. Neuron 56, 14–18.
O'Doherty, J.P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R.J., 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304, 452–454.
Ostrom, E., 2000. Collective action and the evolution of social norms. J. Econ. Perspect. 14, 137–158.

Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 1281–1302 (Dec 1).
Padoa-Schioppa, C., 2007. Orbitofrontal cortex and the computation of economic value. Ann. N. Y. Acad. Sci. 1121, 232–253.
Rangel, A., Camerer, C., Montague, P.R., 2008. A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. 9, 545–556.
Riedl, K., Jensen, K., Call, J., Tomasello, M., 2012. No third-party punishment in chimpanzees. Proc. Natl. Acad. Sci. 109, 14824–14829.
Sanfey, A.G., Rilling, J.K., Aronson, J., Nystrom, L.E., Cohen, J., 2003. The neural basis of economic decision-making in the ultimatum game. Science 300, 1755–1758.
Saxe, R., 2006. Uniquely human social cognition. Curr. Opin. Neurobiol. 16, 235–239.
Singer, T., Seymour, B., O'Doherty, J.P., Kaube, H., Dolan, R.J., Frith, C.D., 2004. Empathy for pain involves the affective but not sensory components of pain. Science 303, 1157–1162.
Spitzer, M., Fischbacher, U., Herrnberger, B., Gron, G., Fehr, E., 2007. The neural signature of social norm compliance. Neuron 56, 185–196.
Tabibnia, G., Satpute, A.B., Lieberman, M.D., 2008. The sunny side of fairness preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychol. Sci. 19, 339–347.
Tricomi, E., Rangel, A., Camerer, C.F., O'Doherty, J.P., 2010. Neural evidence for inequality-averse social preferences. Nature 463, 1089–1091.
Xiang, T., Lohrenz, T., Montague, P.R., 2013. Computational substrates of norms and their violations during social exchange. J. Neurosci. 33, 1099–1108.
Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., Miura, A., Inukai, K., Takagishi, H., Simunovic, D., 2012. Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. Proc. Natl. Acad. Sci. U. S. A. 109, 20364–20368.
Young, L., Camprodon, J.A., 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.
Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., Maier, W., Shah, N.J., Fink, G.R., Zilles, K., 2001 (Jul 31). Mind reading: neural mechanisms of theory of mind and self-perspective. Neuroimage 14 (1), 170–781.
Zaki, J., Mitchell, J.P., 2011. Equitable decision making is associated with neural markers of intrinsic value. Proc. Natl. Acad. Sci. 108, 19761–19766.
Zaki, J., López, G., Mitchell, J.P., 2013. Activity in ventromedial prefrontal cortex covaries with revealed social preferences: evidence for person-invariant value. Soc. Cogn. Affect. Neurosci. nst005.
Zhu, L., Mathewson, K.E., Hsu, M., 2012. Dissociable neural representations of reinforcement and belief prediction errors underlying strategic learning. Proc. Natl. Acad. Sci. U. S. A. 109, 1419–1424.