ARTICLE IN PRESS

YNIMG-12899; No. of pages: 9; 4C: 3, 5, 6, 7

NeuroImage xxx (2016) xxx–xxx

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] School of Business, National University of Singapore, Singapore 119245, Singapore
[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 xxxx

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 author.
E-mail addresses: ecscsh@nus.edu.sg, mhsu@haas.berkeley.edu (M. Hsu).

http://dx.doi.org/10.1016/j.neuroimage.2016.01.040
1053-8119/© 2016 Published by Elsevier Inc.

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 portion 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;