Journal Club

Editor's Note: These short, critical reviews of recent papers in the *Journal*, written exclusively by graduate students or postdoctoral fellows, are intended to summarize the important findings of the paper and provide additional insight and commentary. For more information on the format and purpose of the Journal Club, please see http://www.jneurosci.org/misc/ifa_features.shtml.

Dissociating Guilt- and Inequity-Aversion in Cooperation and Norm Compliance

H Y ,[1] B S ,[1] Y Y ,[1] P R. B ,[1] L J. C [2]

[1]Center for Brain and Cognitive Sciences and Department of Psychology, Peking University, Beijing 100871, China, and [2]Institute for Cognitive Science, Department of Psychology and Neuroscience, University of Colorado, Boulder, Colorado 80309

Review of Nihonsugi et al.

Social norms provide a set of expectations regarding context-specific appropriate behavior that aids in navigating social environments (Bicchieri, 2006). Classic studies have demonstrated that people tend to conform to these norms even at the cost of their own interest (Fehr and Fischbacher, 2004). Expectations vary widely across cultures (Henrich et al., 2001) and there are likely differing motivations for individuals to comply with these norms. For example, one motivation, consequentialism, emphasizes the outcome of an action as the sole measure of its moral worth (Mill, 1861/1998). From this philosophical perspective, one may avoid violating social norms simply because unfair and inequitable outcome are bad for the greater good (e.g., distributional preferences). Alternatively, according to sentimentalism (Smith, 1759/2002), empathy with others "constitutes the moral approval. . . for agents and/or their actions" (Slote, 2010). This framework argues that people are motivated to comply with norms to avoid suffering from harming another as a result of violating the norms (e.g., guilt-aversion).

In reality, these two motivations are likely complementary and each may independently contribute to social decisions with their relative weights varying across individuals and contexts. Unfortunately, the majority of the research that uses social bargaining games to study social decision-making has been unable to effectively dissociate these two distinct motivations. This is likely a consequence of a peculiar convention in bargaining experiments to neither measure nor manipulate individuals' expectations. Thus, it has been unclear how much participants are motivated by distributional preferences (i.e., inequity-aversion) compared with disappointing a relationship partner (i.e., guilt-aversion). Fortunately, there has recently been a growing trend to both measure (Chang et al., 2011; Chang and Sanfey, 2013) and manipulate (Xiang et al., 2013) agents' expectations.

In a recent study published in *The Journal of Neuroscience*, Nihonsugi et al. (2015) provided an important theoretical advance to dissociate the inequity- and guilt-aversion motivations in human norm compliance and identify the brain bases for each motivation. The experimenters used a modified trust game (Charness and Dufwenberg, 2006) in which participants initially decided as an investor whether or not to invest their endowment with an anonymous trustee and reported their belief about the likelihood of the trustee reciprocating. Participants then played the role of the trustee with multiple anonymous investors while undergoing fMRI. For each trial, trustees were given information about the investor's expectation and also the payoffs each player would receive based on their decision to cooperate or defect. For example, if the trustee chose Cooperate, then the investor might receive ¥780 and the trustee ¥650; if the trustee chose Defect, then the investor could receive ¥220 and the trustee ¥910. Though the actual investors' expectations and decisions were predetermined by the experimenters, the trustees were led to believe that they were playing with real agents and were paid proportional to their payoffs in the game at the end of the experiment.

Participants' motivations in the game were inferred based on how much they considered their partners' expectations (e.g., guilt-aversion) and discrepancies between each player's payoffs (e.g., inequity-aversion) when making their decision to cooperate or defect. The basic framework for how these motivations were modeled was based on expected utility theory, which assumes that participants make decisions that maximize their expected payoff. Here, payoffs could be material (based on the amount of money the trustee receives) or psychological (based on concern for the investor's welfare) (Fehr and Camerer, 2007). The authors specifically compared psychological payoffs arising from inequitable distributional outcomes (i.e., the absolute differ-

Received March 29, 2015; revised April 30, 2015; accepted May 5, 2015.

H.Y, B.S, Y.Y, and P.R.B. are supported by grants awarded to Prof. Xiaolin Zhou from the Natural Science Foundation of China (Grants 91232708, 31170972) and the National Basic Research Program of China (973 Program: 2010CB833904, 2015CB856400).

We thank Professor Xiaolin Zhou for his helpful comments on our manuscript. We are also very grateful to Mr. Yin Wu and Mr. Shaorong Yan, without whose support this article would not have been possible.

Correspondence should be addressed to Hongbo Yu, Department of Psychology, Peking University, Beijing 100871, China. E-mail: hbyu101325@pku.edu.cn or abraham.yule@gmail.com.

DOI:10.1523/JNEUROSCI.1225-15.2015

 0270-6474/15/358973-03$15.00/0

ence between the two players' payoffs) (Fehr and Schmidt, 1999) and feelings of guilt, which arose from disappointing a relationship partner by making a decision that resulted in the investor receiving a smaller payoff than he/she expected (i.e., the amount of money that the investor would have received had the trustee chosen to cooperate multiplied by the investor's estimated probability of the trustee's cooperation) (Battigalli and Dufwenberg, 2007). It is important to note that trustees had full information about the investor's expectations and each player's payoffs and thus their motivations can be inferred by how much they considered inequity or disappointing the investor when making their decision. A critical aspect of the experimental design was that the payoff matrix was constructed in such a way that the trustees' payoffs were uncorrelated with the amount of inequity between their partner's payoff, and both were uncorrelated with the investors' expectations about the likelihood the trustee would choose Cooperate. This allowed the experimenters to extend previous work (Chang et al., 2011) and disentangle these two otherwise intertwined motivations underlying human cooperation and norm compliance.

The authors found that the two motivations were associated with different neural circuitry. Controlling for guilt, inequity was positively associated with activation in the ventral striatum and amygdala. While other studies have implicated the ventral striatum in tracking inequity, it appears to go in the opposite direction, such that there is greater ventral striatal and amygdala activation associated with decreasing inequity (Tabibnia et al., 2008; Tricomi et al., 2010). There are several possible reasons that can account for these discrepancies. First, these studies differed substantially in their design. In this study, the participants made decisions based on the inequity of the payoffs, while participants in the Tricomi et al. (2010) passively observed inequitable divisions and participants in the Tabibnia et al. (2008) made decisions to punish based on the unfairness of the offer. Second, unlike the other studies, Nihonsugi et al. (2015) have experimentally and statistically controlled for guilt.

In addition, after controlling for inequity, the authors found that guilt-aversion was positively associated with activity of the right dorsolateral prefrontal cortex (rDLPFC). This region, among others (such as the anterior cingulate cortex and insula), has been associated with guilt in both decision-making and error-monitoring paradigms (Chang et al., 2011; Koban et al., 2013). Though it is important to note that guilt-aversion may not necessarily be equivalent to the feelings of guilt that result from explicitly knowing that one's actions resulted in harm to another (Koban et al., 2013; Yu et al., 2014), as participants make decisions that minimize their anticipated guilt. An additional strength of the Nihonsugi et al. (2015) study is that the authors followed up their rDLPFC finding and went on to test the causal role of this area in such processing. Utilizing a noninvasive brain stimulation technique known as transcranial direct current stimulation (tDCS), the authors used anodal stimulation of the rDLPFC while the participants played the trust game. Anodal stimulation is thought to temporarily enhance the neuronal excitability of cortex and has been associated with increasing BOLD activation in primary motor cortex (Stagg et al., 2009). Impressively, the authors demonstrated that participants showed more reliance on guilt-aversion when making their decisions to honor trust when the rDLPFC function was enhanced relative to a sham stimulation, whereas the weight of inequity appeared to be unaffected by the anodal stimulation. Together, these results provide compelling evidence implicating the rDLPFC in processing sentiment-based (i.e., guilt-aversion), but not the outcome-based (i.e., inequity-related) motivations.

It is interesting to consider the current finding in light of another recent investigation of the role of the rDLPFC in norm compliance (Ruff et al., 2013). In that study, the authors used both anodal and cathodal (decreasing neuronal excitability) stimulation of the rDLPFC while players decided how much money to share with a partner. The authors found that anodal stimulation increased contributions when there was a possibility of a sanction, but decreased contributions when there was no threat of sanction. One interpretation of Ruff et al.'s (2013) results, consistent with the findings reported by Nihonsugi et al. (2015), is that each condition was associated with different norms. For example, the descriptive norm for contributions in the context of a threat of sanction will likely be higher than in the no sanction condition. If the rDLPFC is associated with motivating behavior to minimize guilt-aversion and increase norm-compliance, then, based on Nihonsugi et al.'s (2015) findings, anodal stimulation should increase contributions when the norm is high and paradoxically decrease contributions when the norm is low (Sanfey et al., 2014). In other words, by increasing the consideration of others' expectations, the rDLPFC enhancement shifts the participants' behavior to align with such expectations (Nihonsugi et al., 2015).

There are a number of promising extensions. First, in the current study, the investor's belief about the likelihood of the trustee's cooperation was explicitly presented to the trustee. However, in real social interactions, people use mentalizing to infer what their partner expects. Future work could more explicitly model this process using techniques such as Bayesian modeling (Yoshida et al., 2008). Second, this study nicely demonstrates how guilt-aversion affects decision-making. Future work might examine how such expectations modulate the experience (Yu et al., 2014), rather than just the anticipation of guilt. Combining computational modeling approaches with innovative experimental designs is a promising avenue for uncovering the neural processes involved in social cognition and decision-making.

One important caveat when interpreting Nihonsugi et al.'s (2015) results is that it remains unclear how much of their effect is purely related to guilt-aversion and how much can be attributed to reputational concerns. For example, the trustee might not actually feel guilty by the prospect of disappointing a partner, but simply does not want to be regarded poorly in a social situation by both the investor and the third-party experimenter. The original Charness and Dufwenberg (2006) experiment had a random chance element following the trustee's decision, which specifically provided a mechanism to rule out the reputational concern. Indeed, a previous study has demonstrated that disrupting the rDLPFC with transcranial magnetic stimulation disrupts trustees' concerns about developing a good reputation in the trust game (Knoch et al., 2009), which is entirely consistent with the findings of the current study. Future work should attempt to carefully experimentally disentangle guilt-aversion from reputational concerns.

Overall, this work marks the strength of the computational approach to complex social behaviors and affective processes, as it successfully dissociated two distinct psychological and neural mechanisms underlying human cooperation and norm compliance (Fehr and Schmidt, 1999; Charness and Dufwenberg, 2006). The authors nicely demonstrate that both

consequentialism and sentimentalism considerations independently affect norm compliance and cooperation. Moreover, these motivations appear to be encoded in separate brain circuits. We believe that combining formal mathematical modeling, neuroscientific techniques, and social psychological theories will continue to further our insight into the material basis of our social nature.

References

Battigalli P, Dufwenberg M (2007) Guilt in games. Am Econ Rev 97:170–176. CrossRef

Bicchieri C (2006) The grammar of society: the nature and dynamics of social norms. Cambridge: Cambridge UP.

Chang LJ, Sanfey AG (2013) Great expectations: neural computations underlying the use of social norms in decision-making. Soc Cogn Affect Neurosci 8:277–284. CrossRef Medline

Chang LJ, Smith A, Dufwenberg M, Sanfey AG (2011) Triangulating the neural, psychological, and economic bases of guilt aversion. Neuron 70:560–572. CrossRef Medline

Charness G, Dufwenberg M (2006) Promises and partnership. Econometrica 74:1579–1601. CrossRef

Fehr E, Camerer CF (2007) Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci 11:419–427. CrossRef Medline

Fehr E, Fischbacher U (2004) Third-party punishment and social norms. Evol Hum Behav 25:63–87. CrossRef

Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. Q J Econ 114:817–868. CrossRef

Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, McElreath R (2001) In search of Homo economicus: behavioral experiments in 15 small-scale societies. Am Econ Rev 91:73–78. CrossRef

Knoch D, Schneider F, Schunk D, Hohmann M, Fehr E (2009) Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. Proc Natl Acad Sci U S A 106:20895–20899. CrossRef Medline

Koban L, Corradi-Dell'Acqua C, Vuilleumier P (2013) Integration of error agency and representation of others' pain in the anterior insula. J Cogn Neurosci 25:258–272. CrossRef Medline

Mill JS (1861/1998) Utilitarianism (Crisp R, ed). New York: Oxford UP.

Nihonsugi T, Ihara A, Haruno M (2015) Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. J Neurosci 35:3412–3419. CrossRef Medline

Ruff CC, Ugazio G, Fehr E (2013) Changing social norm compliance with noninvasive brain stimulation. Science 342:482–484. CrossRef Medline

Sanfey AG, Stallen M, Chang LJ (2014) Norms and expectations in social decision-making. Trends Cogn Sci 18:172–174. CrossRef Medline

Slote MA (2010) Moral sentimentalism. New York: Oxford UP.

Smith A (1759/2002) The theory of moral sentiments (Haakonssen K, ed). Cambridge: Cambridge UP.

Stagg CJ, Best JG, Stephenson MC, O'Shea J, Wylezinska M, Kincses ZT, Morris PG, Matthews PM, Johansen-Berg H (2009) Polarity-sensitive modulation of cortical neurotransmitters by transcranial stimulation. J Neurosci 29:5202–5206. CrossRef Medline

Tabibnia G, Satpute AB, Lieberman MD (2008) The sunny side of fairness preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychol Sci 19:339–347. CrossRef Medline

Tricomi E, Rangel A, Camerer CF, O'Doherty JP (2010) Neural evidence for inequality-averse social preferences. Nature 463:1089–1091. CrossRef Medline

Xiang T, Lohrenz T, Montague PR (2013) Computational substrates of norms and their violations during social exchange. J Neurosci 33:1099–1108. CrossRef Medline

Yoshida W, Dolan RJ, Friston KJ (2008) Game theory of mind. PLoS Comput Biol 4:e1000254. 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