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.

Modeling Emotion and Learning of Norms in Social Interactions

Luke J. Chang and Leonie Koban

Institute for Cognitive Science, Department of Psychology & Neuroscience, University of Colorado, 345 UCB, Boulder, Colorado 80309

Review of Xiang et al.

Social norms are shared expectations about appropriate behavior for a specific context and have been extensively studied in social psychology. Classic experiments demonstrate that we are motivated to behave consistently with "descriptive norms," that is, what we believe most people would do, even when it counters our own beliefs (Asch, 1956). Yet, little is known about the neural processes that determine how we learn socially appropriate behavior for every context, and why we are motivated to conform to these beliefs.

Recently, social norms have been studied in the field of behavioral economics (Fehr and Fischbacher, 2004; Bicchieri, 2006) using social dilemmas such as the Ultimatum Game. In this task, Player 1 is endowed with a sum of money and can propose to divide this money any way he/she chooses with his/her partner. Player 2 ultimately decides whether to accept the proposed monetary offer, or alternatively, he/she can reject the offer, in which case neither player receives any money. Theoretically, if people are solely motivated by monetary gains, then Player 2 should accept any nonzero offer and Player 1 should anticipate this and make the smallest possible offer (e.g., $1). However, experimental evidence indicates that Player 1s typically offer 30–50% of their endowment and Player 2s tend to reject proposals' lower than 20% half of the time (Camerer, 2003). To account for this finding it has been proposed that people may prefer equitable outcomes when interacting with other agents (Fehr and Fischbacher, 2004). Although this theory has been widely adopted as an explanation for cooperative behavior, it has been relatively unsuccessful in explaining behavior in diverse samples with varying values and cultural norms (Henrich et al., 2006). Therefore, it has been suggested that both players may use a descriptive norm as a reference point in these social dilemmas and conform their behavior to what they believe most others will do (Bicchieri, 2006; Chang and Sanfey, 2011). The neural signal associated with the detection of social conflict and subsequent conforming to descriptive norms appears to originate in the dorsal anterior cingulate cortex (ACC) and ventral striatum (Klucharev et al., 2009; Koban et al., 2013a). Recent work directly measuring expectations has found that deviations from expected offers in the Ultimatum Game correlate with activity in the ACC, and ultimately prompt rejection of offers in the game (Chang and Sanfey, 2011). However, how people form and update their beliefs about social norms and whether violations of these beliefs are linked to affective signals that motivate behavioral adaptations remain open questions.

A recent fMRI study by Xiang et al. (2013) sought to examine these questions by manipulating the normative offer encountered in a series of single round Ultimatum Games. Participants were assigned the role of Player 2 and divided into different experimental groups in which they were presented with offers that were drawn from one of three Gaussian distributions (e.g., high, medium, and low offers). This manipulation effectively conditioned the participants' expectations about the type of offers they would encounter in the game. Unbeknownst to the participants, the distributions changed after the first half of the experiment such that the high and low groups received offers from a medium distribution. This allowed the authors to directly compare the effects of expectations on participants' decisions and emotional experiences. Participants who initially encountered high offers rejected lower offers more frequently than participants who initially experienced lower offers. This behavioral effect was potentially emotionally motivated as it was associated with higher ratings of negative affect.

An important aspect of the authors' approach is that they used formal models to understand how previous experience would affect behavior and emotions. Mathematically, descriptive social norms can be formulated as a belief about the likelihood of most people's behavior in a given context. An expectation is simply

Received March 1, 2013; revised April 2, 2013; accepted April 5, 2013.

L.K. is supported by a post-doctoral fellowship (PBGEP1_142252) from the Swiss National Science Foundation. We thank Eunice Lee and Drs. Alec Smith, Elizabeth Losin, and Hedwig Eisenbarth for their helpful comments on our manuscript.

Correspondence should be address to Dr. Luke J. Chang, Institute for Cognitive Science, Department of Psychology & Neuroscience, University of Colorado, 345 UCB, Boulder, Colorado 80309. E-mail: luke.chang@colorado.edu.

DOI:10.1523/JNEUROSCI.0973-13.2013

defined as the mean of this probability distribution. Formalizing models of how social norms are updated and influence decisions provides specific predictions that can be tested behaviorally. These models can also be used to generate regressors in functional imaging analyses to identify brain regions that are presumably involved in these computations.

Xiang et al. (2013) combined three different functions in their modeling approach. First, similar to previous studies, they used a behavioral economic "utility" function to quantify the value associated with accepting each offer. This function posits that the value of a decision depends not only on the amount of money offered, but also on the extent to which the offer deviates from the social norm. Second, the utility associated with each offer was placed into a "softmax" function, which calculates the probability that the responder would accept versus reject an offer. Third, beliefs about the social norm were estimated using an "Ideal Bayesian Observer." This function is the critical innovation in this study, and extends previous work (Fehr and Fischbacher, 2004; Chang and Sanfey, 2013) by describing how beliefs about the social norm change as a function of previous social interactions (i.e., offers in previous trials). The experience-based modeling of norms with an Ideal Bayesian Observer provides the optimal way to integrate new information (e.g., offer in the current trial) with prior beliefs (from previous learning) to form a new belief. In this framework, the "norm prediction error" is defined as the difference between an encountered offer and the prior expectation and reflects the degree to which an offer deviates from a social norm. "Variance prediction error" is defined as the difference between the squared prediction error and the mean of the prior variance distribution and can be understood as prediction errors regarding the uncertainty or variance of a distribution. The authors used trial-to-trial predictions of the norm prediction error and variance prediction error to examine where these computational processes are instantiated in the brain.

The results indicated that norm prediction errors were positively correlated with activity in the ventral striatum and ventromedial prefrontal cortex (vmPFC), which is consistent with previous work that has outlined where prediction error is computed in the brain using simple reward learning tasks (O'Doherty et al., 2003). In contrast, the variance prediction error was associated with activity in the insula and ACC, consistent with the hypothesis that these regions process unexpected outcomes, regardless of whether they are better or worse than expected (Preuschoff et al., 2008). Together, these findings provide compelling support that the learning of social norms recruits the same neural systems involved in basic learning processes.

An important additional implication of the Xiang et al. (2013) study is that violations of social norms may create an affective error signal, which motivates us to conform to and enforce social norms (Montague and Lohrenz, 2007). Previous work has suggested that social emotions, such as guilt, arise when subjects deviate from social expectations and that "social error" signals may serve as the primary motivation to conform or adapt behavior (Chang et al., 2011; Koban et al., 2013b). Alternatively, observing others violating social norms may create a different affective response such as anger, which motivates enforcement of the social norm (Chang and Sanfey, 2013). Consistent with this hypothesis, Xiang et al. find that norm prediction errors explained a substantial amount of the variance (38%) of participants' affective ratings. In addition, affective prediction errors were associated with activity in the ventral striatum and vmPFC, while affective variance prediction errors were associated with activation in the anterior insula. These results parallel the neural correlates of the norm and variance prediction errors and suggest that affective and social norm prediction error may be reflecting the same process.

While this work provides an important step in characterizing the neural computations associated with social norms, there are a number of promising areas for future work. First, the extent to which individuals learn according to an ideal Bayesian learner remains unclear. Surely, there is individual variability in how people learn from positive and negative social feedback. Second, while this work favors a descriptive social norm account compared with inequity aversion theory, it would be interesting to model additional motivations in the utility function such as others' intentions or individual moral values, which could help to generalize this utility function to other contexts. Third, it would be interesting to investigate Player 1's affective and neural responses to expected and unexpected rejections of his/her offers by Player 2 in this dyadic game. These neural signals may provide a feedback signal by which he/she updates beliefs about the norm (Fareri et al., 2012). Finally, while the authors nicely demonstrate a link between norm and variance prediction error and emotions, future work still needs to demonstrate that emotions motivate conformity to social norms and determine whether the prediction error is sufficient to account for the cognitive aspects of the emotion.

Overall, this work provides a promising development for social and emotion research as it illustrates how the computations associated with social norms can be formalized with mathematical functions and used to understand how these processes are instantiated in the brain. This research contributes to a growing interest in understanding the neural computations associated with social cognition such as mentalizing (Hampton et al., 2008) and trusting advice (Behrens et al., 2008), as well as social emotions such as guilt (Chang et al., 2011; Koban et al., 2013b). In addition, these computational substrates will likely prove to be useful phenotypes for characterizing the social and affective deficits associated with psychopathology (Montague et al., 2012). While these computational approaches have primarily been developed in the context of learning and decision-making, they can readily be extended to other domains of social and affective neuroscience. We believe that adopting a more formal framework will be invaluable in propelling social neuroscience research by providing a mathematical architecture to test specific computational hypotheses and also evaluate competing hypotheses.

References

Asch SE (1956) Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychological Monographs 70:70.

Behrens TE, Hunt LT, Woolrich MW, Rushworth MF (2008) Associative learning of social value. Nature 456:245–249. CrossRef Medline

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

Camerer CF (2003) Behavioral game theory. New York, NY: Russell Sage Foundation.

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

Fareri DS, Chang LJ, Delgado MR (2012) Effects of direct social experience on trust decisions and neural reward circuitry. Front Neurosci 6:148. CrossRef Medline

Fehr E, Fischbacher U (2004) Social norms and human cooperation. Trends Cogn Sci 8: 185–190. CrossRef Medline

Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci U S A 105:6741–6746. CrossRef Medline

Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, Bolyanatz A, Cardenas JC, Gurven M, Gwako E, Henrich N, Lesorogol C, Marlowe F, Tracer D, Ziker J (2006) Costly punishment across human societies. Science 312:1767–1770. CrossRef Medline

Klucharev V, Hytönen K, Rijpkema M, Smidts A, Fernández G (2009) Reinforcement learning signal predicts social conformity. Neuron 61: 140–151. CrossRef Medline

Koban L, Pichon S, Vuilleumier P (2013a) Responses of medial and ventrolateral prefrontal cortex to interpersonal conflict for resources. Soc Cogn Affect Neurosci. Advance online publication. Retrieved April 9, 2013. Medline

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

Montague PR, Lohrenz T (2007) To detect and correct: norm violations and their enforcement. Neuron 56:14–18. CrossRef Medline

Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends Cogn Sci 16:72–80. CrossRef Medline

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38:329–337. CrossRef Medline

Preuschoff K, Quartz SR, Bossaerts P (2008) Human insula activation reflects risk prediction errors as well as risk. J Neurosci 28:2745–2752. 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