Cell PRESS
Neuron
Minireview

To Detect and Correct: Norm Violations and Their Enforcement

P. Read Montague[1,*] and Terry Lohrenz[1]
[1]Department of Neuroscience, Computational Psychiatry Unit, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA
*Correspondence: read@bcm.tmc.edu
DOI 10.1016/j.neuron.2007.09.020

Compliance with social norms requires neural signals related both to the norm and to deviations from it. Recent work using economic games between two interacting subjects has uncovered brain responses related to norm compliance and to an individual's strategic outlook during the exchange. These brain responses possess a provocative relationship to those associated with negative emotional outcomes, and hint at computational depictions of emotion processing.

> *Life's unfairness is not irrevocable; we can help balance the scales for others, if not always for ourselves.*
>
> —Hubert Humphrey

The late Hubert Humphrey was making a political appeal in this quote, but the same language highlights one important feature of our social instincts—humans have an automatic drive to "balance the scales" for social wrongs perpetrated on themselves and others. A large bully takes the lunch money of a smaller child; a huckster sells defective goods to a mentally challenged person, and so on. Such scenarios can motivate us to act on behalf of the wronged person—whether we do or not is another issue. But it's the instinct that's most interesting because it displays our deep connection to others—basically, we want to *balance the other person's ledger*. Social debts to you have representation in my nervous system in such a way that I may be willing, at a cost to myself, to seek some kind of repayment from the bad actor that wronged you. It's an old story chronicled throughout history, but it suggests that our nervous system values the experiences of other individuals, compares them to expected norms, and generates the desire to act if the other person's experience differs too much from the norm.

Modeling Others

The capacity to richly model other agents should be expected in any sufficiently complex social creature that must estimate the intentions and actions of others. But to sense a "social debt" requires specific computational substrates: (1) a shared norm about what is expected, (2) the capacity to detect ongoing deviations from that norm, and (3) the capacity to do this from a third-person perspective, that is, from the point-of-view of the offended individual. Without these basic capacities, we would not reasonably expect a creature to be able to even sense norm violations, much less care about them when they happen to others. This list is obviously not complete, but a creature possessed of these abilities can better guess the likely internal states and likely actions of others and consequently can make better decisions in the context of others.

But here's the interesting point. Our nervous systems do more than simply model the recipient of a social offense; our nervous systems "care" about these offenses to the extent that we are motivated to "right the wrong" for the other person. To be motivated to balance the ledger for an unrelated individual and to possess neural responses related to such drives highlights the importance of uncovering the neural underpinnings of fairness instincts, altruism, and the many other social sentiments that fall into this category (Rilling et al., 2002; Sanfey et al., 2003; de Quervain et al., 2004; King-Casas et al., 2005; Delgado et al., 2005; Singer et al., 2006; Spitzer et al., 2007 [this issue of *Neuron*]).

Given these observations, how can one probe fairness instincts in humans using neuroimaging? In recent years, this genre of question has been addressed using economic exchange games in combination with either PET scans or functional magnetic resonance imaging (fMRI). The games have a variety of names—the well-known prisoner's dilemma, the dictator game, the ultimatum game, the trust game, and so on (Axelrod, 1984; Guth et al., 1982; Roth, 1995; Camerer, 2003). They are excellent experimental probes because they are simple and mathematically well-specified, there is an existing body of behavioral data employing them across a variety of contexts, and there are known solution concepts for how they "should" be played by a rational self-interested agent (Roth, 1995; Camerer, 2003; Camerer and Fehr, 2006). Most importantly, they all require participants to model their partner.

Fairness Games Expose Norms and Their Error Signals

OK, so the economic games are quantitatively prescribed and possess known optimal or near-optimal solutions. But the really important part of their structure is the requirement to model the other player. To do so, players must share reasonably similar norms for what constitutes an expected behavior, and this requirement holds *before* any

economic exchange occurs. Take the ultimatum game as an example.

The ultimatum game involves two players—the proposer and the responder (Guth et al., 1982), and could reasonably be renamed "take-it-or-leave-it." In this game, the proposer is endowed with some resource (say $100) and can offer any split to the responder. Let's suppose the proposer offers $80 for herself and $20 for the responder. If the responder accepts the split, then both players walk away with money ("take it" option). If the responder rejects, neither player gets anything ("leave it" option). Rationally, the responder should accept all nonzero offers since they start with nothing, but experiments show this expectation to be false. In practice, the proposer sends $40 as their modal offer and responders reject 50% of the time at an $80:$20 split.

The neural responses engendered by this game are provocative because of their relationship to neural responses measured during negative emotional events. In fairness games like the ultimatum game, at the revelation of the proposer's offer, the anterior insula of the responder's brain activates parametrically to offer level and by doing so correlates with the degree of unfairness in the offered split (Sanfey et al., 2003). This response also covaries with the probability that the responder will reject the offer, thus providing a neural signature for the likelihood of punishing the proposer at the responder's expense. The importance of the responses to this fairness game derives from the vast array of other task demands that also activate this same region of the insula. These include hunger, thirst, anger, moral and physical disgust, sadness (induced by scripted stories), dread of future shock, anticipation of emotionally aversive events, pain, a host of other interoceptive events, and the list goes on (e.g., Phillips et al., 1997; Ploghaus et al., 1999; Damasio et al., 2000; Critchley et al., 2004; Craig, 2002, 2005; Berns et al., 2006; Stein et al., 2007). Lesions in this region have even been linked with a decrease in the ability of smoking cues to elicit craving in smokers (Naqvi et al., 2007; Dani and Montague, 2007; Gray and Critchley, 2007). A question naturally arises: what kind of computation would respond similarly to such a spectrum of stimuli; one that ranges from physical disgust to unfairness in a monetary exchange game?

We think that a computational depiction of these data is warranted and sheds light on a new study from Fehr and colleagues in this issue of *Neuron* (Spitzer et al., 2007). The idea is motivated by a variety of findings, but a recent result by Preuschoff et al. (2006) (Figure 1) using an economic task points the way to a model, which we sketch in Figure 2. In this experiment (Figure 1A), the aim was to track separately hemodynamic responses to the expected value and variance of a future reward (money). This experiment showed that well-known dopaminoceptive structures in the striatum possess ongoing hemodynamic signatures related to both expected value and variance of the future payoff (Preuschoff et al., 2006). But another important finding was that bilateral anterior insula also showed an ongoing hemodynamic response to the variance in payoff (i.e., risk) and to an error signal related to this variance (also see Preuschoff and Bossaerts, 2007; Knutson and Bossaerts, 2007).

Connecting Norm Error Signals to a Range of Emotions

If we back up slightly from the specifics of this financial task, we can combine these observations with those of the ultimatum game (Sanfey et al., 2003) that show anterior insula sealing with the degree of unfairness of the proposer's offer, that is, how much the offer differs from some shared norm across the two players. These experiments point the way to a broader view of the computations available at the level of the anterior insula. The simplest one-dimensional version of the idea is that a norm along some behavioral dimension is being compared to an ongoing estimate along that same dimension derived from the creature's actual experience (red and green curves in Figure 2). The game to play is this: the creature should want to express a behavior and/or changes in its internal states to *match the estimated distribution to the norm*. The technical method used to do such matching is not important here, but we note that natural "error signals" for matching one distribution to another are errors in the first two moments of the distributions, that is, errors in the means and variances of the distributions. We also leave open the issue of whether the animal seeks to move the estimate to match the norm or moves the norm to match the estimate or both.

As noted above, a wide range of negative emotional events activate this same brain region as well as changes or forecast changes in interoceptive states (Paulus et al., 2003; Critchley et al., 2004; see Craig, 2002, 2005 for reviews). This suggests the hypothesis that these emotions are underwritten by error signals along some norm distribution, suggesting that emotions like disgust, pain, thirst, anger, and so on represent feelings associated with their own unique error signals used to direct the organism to "zero" them by changing its behavior and/or internal states. This class of idea has been bruiting about the literature in different forms but less committed to a computational interpretation (Critchley et al., 2004; Gray and Critchley, 2007; Stein et al., 2007). However, we suggest that the economic game results recommend a more general situation where each emotional state is associated with specific norm error signals that guide the matching of some estimated distribution to the stored norm.

Matching a Norm through the Threat of Enforcement

These ideas will mature in the coming years, but they point the way to the current paper by Spitzer et al. (2007). This paper examines the behavioral and neural correlates of norm compliance. To comply with any norm, a nervous system must possess a representation of the norm and a capacity to decide whether the norm is being met, that is, error signals related to the norm. Compliance with

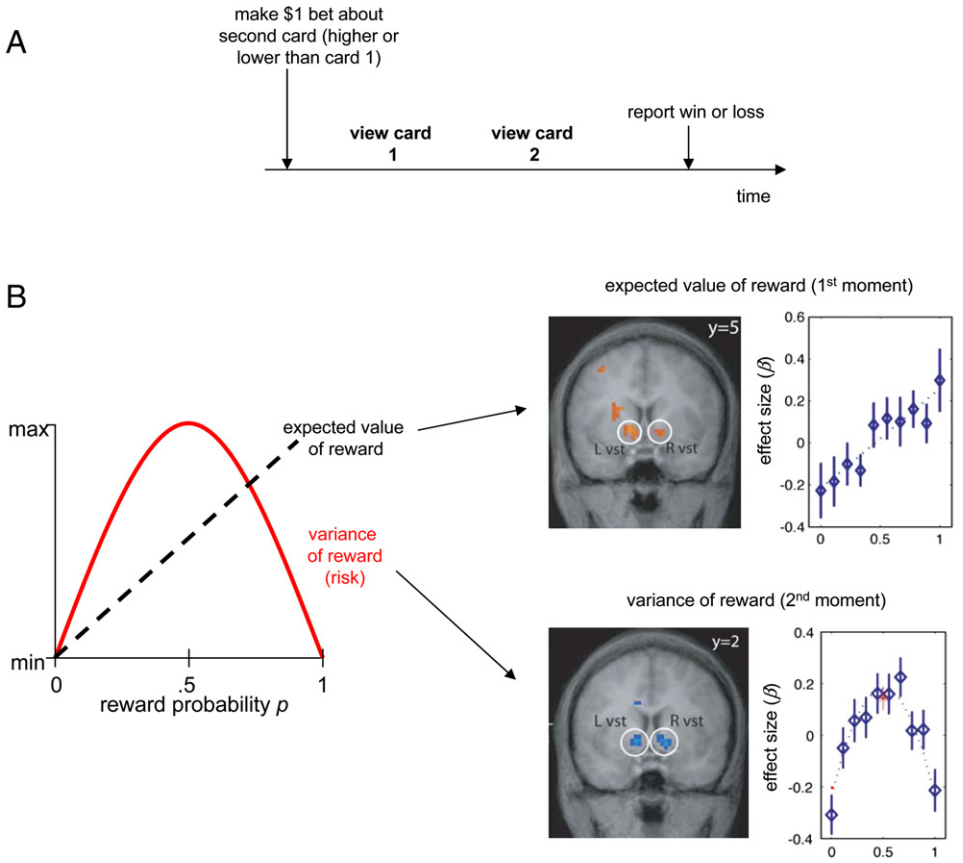

Figure 1. Neural Responses to Expected Value and Variance of Future Reward

(A) Timeline of events for two-card task. Subjects place a bet about whether a second card will be higher or lower than the first card. There are only ten cards and this bet is placed before seeing either card. At that time, the probability of reward is 1/2.

(B) After the first card is shown, the expected value of reward scales linearly with probability of reward p while the variance scales quadratically with p. The two insets show activations in ventral striatum to expected value of payoff (top, orange activations) and variance of payoff (bottom, blue activations). The insets show plots of expected value of reward versus probability (top) and variance of reward versus probability (bottom) both for the left ventral striatum.

Adapted from Preuschoff et al. (2006).

social norms is really never quite voluntary. An individual can be compelled by some internal norm and guided by natural error signals related to it (e.g., guilt). Alternatively, an individual can also be coerced through signals from other agents like institutions or individuals around them (Fehr and Gachter, 2002). These signals and their relationship to our societal norms form one of the most important and interesting parts of human cognitive activity, but one for which detailed neural data has been sorely lacking (but see de Quervain et al., 2004).

In their paper, Spitzer et al. (2007) look at the behavior and brain activity of subjects playing a variant of the dictator game in two separate conditions (Figure 3). In both conditions, each player is endowed with 25 monetary units (mu's). In the *no-punishment condition*, the first player (A) decides how to split an additional 100 mu's between herself and player B. In this condition, player B has no chance to punish. In the *punishment condition*, player B can "buy" punishment units in a 5:1 ratio. For example, if player A were to keep the entire 100, player B could spend her 25 mu's to wipe out player A's entire stake.

Not surprisingly, subjects in this experiment transferred on average more mu's in the punishment condition than in the control condition. During subject A's decision period, and contrasting punishment versus control condition, Spitzer et al. (2007) found increased activation in the dorsolateral prefrontal cortex (DLPFC), ventrolateral prefrontal cortex (VLPFC), bilateral orbitolateral prefrontal cortex (OLPFC), and caudate nucleus.

Two findings stand out in this study. Brain responses in OFC, DLPFC, and caudate correlated significantly with changes in amount transferred across conditions—here, differences in the amount transferred between *punishment possible* and *punishment not possible* conditions served as the monetary measure of norm compliance. OLPFC has previously been implicated in the evaluation of negative stimuli, while DLPFC activates during tasks requiring cognitive control (Miller and Cohen, 2001). The second standout finding was the response that correlated

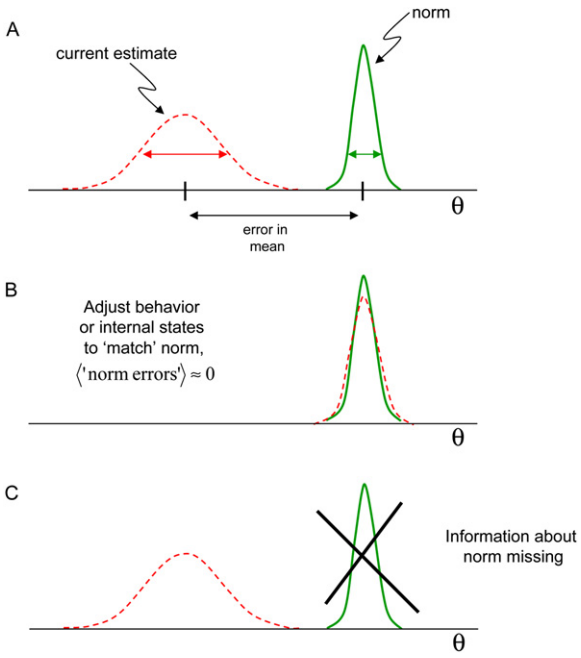

Figure 2. Norm Errors Provide Natural Computational Substrate for Emotions

(A) To detect and correct norm violations there must be a representation of the norm along the relevant dimension(s) (green curve). To compare the creature's current estimate of its state to the norm, some mechanism must at least be able to sense and correct the difference in the means and variances in the distributions.

(B) By changing its behavior and internal states a creature can match its estimated distribution along θ to the norm for that dimension. The exact mathematical mechanism for how this is accomplished is not important here, but note that either the estimate and/or the norm might change to effect this matching.

(C) In this panel, the capacity to represent the norm has been lost, and so the creature's nervous system cannot generate errors related to the mismatch between its estimate and the norm. In this case, one might also expect no "corrective" feelings associated with the missing norm. This case is reminiscent of recent work by Naqvi et al. (2007) (see also Dani and Montague, 2007) where lesions to anterior insular cortex caused subjects to "forget to want to smoke." For some reason, the missing insular cortex correlated with an inability of smoking cues to elicit error signals related to craving (Gray and Critchley, 2007). This account would show that minimally such subjects would not have norm error signals necessary for such feelings.

with each subject's strategic nature as measured by their Machiavellian score. The Machiavelli score is based on a questionnaire and asks subjects if they agree with statements such as "It is hard to get ahead without cutting corners here and there." There was a strong correlation in the lateral OFC and insular cortex between differential activation across punishment conditions and the Machiavelli scores. This finding plugs right into our preceding discussion about the computations available at the level of the insular cortex. By definition, a strong Machiavellian impulse deviates significantly from the norm.

One may rightly ask if the social context matters in the punishment condition. The authors conducted an additional control experiment in which the player A played

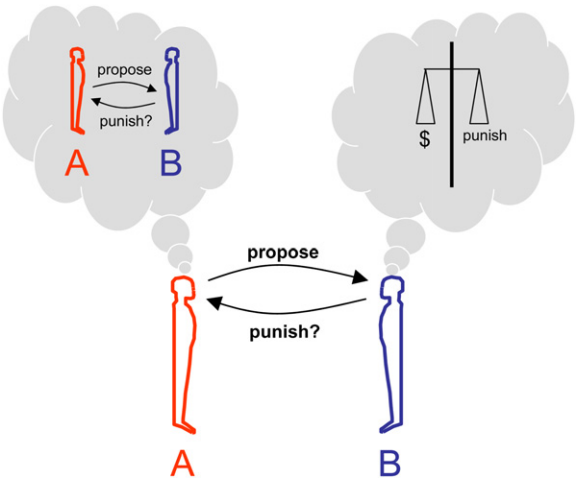

Figure 3. Economic Exchange Tasks Reveal Neural Responses to Models of Others

Here, we depict the dictator game played with two subjects and the possibility of punishment. In the standard dictator game, player A, the dictator, is endowed with some amount of money and can choose a split of that money with player B using only their sense of fairness as their guide. This and related games provide a simple class of behavioral probes for how each player models the other with only a few variables to consider. If subject B is given the option of penalizing or punishing player A for too small a transfer, then subject A will send more. This game and others like it show how each player possesses a norm for what is fair and also possesses a good model of how the other player will react if those norms are violated. These games may help to uncover computational learning signals related to emotional processing.

a computer (and were told they were playing a computer) programmed to punish with the same distribution as human subjects. In the regions activated in the punishment versus no-punishment conditions, the R DLPFC and R OLPFC exhibited significantly more activation in the social context. Whether this is a simple arousal effect ("people are more important as opponents") remains to be seen. This study extends significantly our understanding of the neural processes involved in decision making during active social interactions. In particular, this study shows that areas recruited in evaluating threat and exercising control are employed in this task at a level beyond that in a similar interaction with a computer. More intriguingly, this study shows that there is a very specific functional neural difference between high-Machiavellian and low-Machiavellian types. I guess we shouldn't be surprised.

REFERENCES

Axelrod, R.M. (1984). The Evolution of Cooperation (New York: Basic Books).

Berns, G.S., Chappelow, J., Cekic, M., Zink, C.F., Pagnoni, G., and Martin-Skurski, M.E. (2006). Science *312*, 754–758.

Camerer, C.F. (2003). Behavioral Game Theory (Princeton: Princeton University Press).

Camerer, C.F., and Fehr, E. (2006). Science *311*, 47–52.

Craig, A.D. (2002). Nat. Rev. Neurosci. *3*, 655–666.

Craig, A.D. (2005). Trends Cogn. Sci. *9*, 566–571.

Critchley, H.D., Wiens, S., Rotshtein, P., Ohman, A., and Dolan, R.J. (2004). Nat. Neurosci. *7*, 189–195.

Damasio, A.R., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L., Parvizi, J., and Hichwa, R.D. (2000). Nat. Neurosci. *3*, 1049–1056.

Dani, J.A., and Montague, P.R. (2007). Nat. Neurosci. *10*, 403–404.

Delgado, M.R., Frank, R.H., and Phelps, E.A. (2005). Nat. Neurosci. *8*, 1611–1618.

de Quervain, D.J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004). Science *305*, 1254–1258.

Fehr, E., and Gachter, S. (2002). Nature *415*, 137–140.

Guth, W., Schmittberger, R., and Schwarze, B. (1982). J. Econ. Behav. Organ. *3*, 367–388.

Gray, M.A., and Critchley, H.D. (2007). Neuron *54*, 183–186.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., and Montague, P.R. (2005). Science *308*, 78–83.

Knutson, B., and Bossaerts, P. (2007). J. Neurosci. *27*, 8174–8177.

Miller, E.K., and Cohen, J.D. (2001). Annu. Rev. Neurosci. *24*, 167–202.

Naqvi, N.H., Rudrauf, D., Damasio, H., and Bechara, A. (2007). Science *315*, 531–534.

Paulus, M.P., Rogalsky, C., Simmons, A., Feinstein, J.S., and Stein, M.B. (2003). Neuroimage *19*, 1439–1488.

Phillips, M.L., Young, A.W., Senior, C., Brammer, M., Andrew, C., Calder, A.J., Bullmore, E.T., Perrett, D.I., Rowland, D., Williams, S.C., et al. (1997). Nature *389*, 495–498.

Ploghaus, A., Tracey, I., Gati, J.S., Clare, S., Menon, R.S., Matthews, P.M., and Rawlins, J.N. (1999). Science *284*, 1979–1981.

Preuschoff, K., and Bossaerts, P. (2007). Ann. N Y Acad. Sci. *1104*, 135–146.

Preuschoff, K., Bossaerts, P., and Quartz, S.R. (2006). Neuron *51*, 381–390.

Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., and Kilts, C.D. (2002). Neuron *35*, 395–405.

Roth, A. (1995). In Handbook of Experimental Economics, J.H. Kagel and A.E. Roth, eds. (Princeton, NJ: Princeton Univ. Press), pp. 253–348.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2003). Science *300*, 1755–1758.

Stein, M.B., Simmons, A., Feinstein, J.S., and Paulus, M.P. (2007). Am. J. Psychiatry *164*, 318–327.

Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., and Frith, C.D. (2006). Nature *439*, 466–469.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., and Fehr, E. (2007). Neuron *56*, this issue, 185–196.