Behavioral/Cognitive

Computational Substrates of Norms and Their Violations during Social Exchange

Ting Xiang,[1]* Terry Lohrenz,[2]* and P. Read Montague[2,3]

[1]Department of Neuroscience, Baylor College of Medicine, Houston, Texas 77030, [2]Virginia Tech Carilion Research Institute and Department of Physics, Virginia Tech, Roanoke, Virginia 24016, and [3]Wellcome Trust Centre for Neuroimaging, London, WC1N 3BG, United Kingdom

Social norms in humans constrain individual behaviors to establish shared expectations within a social group. Previous work has probed social norm violations and the feelings that such violations engender; however, a computational rendering of the underlying neural and emotional responses has been lacking. We probed norm violations using a two-party, repeated fairness game (ultimatum game) where proposers offer a split of a monetary resource to a responder who either accepts or rejects the offer. Using a norm-training paradigm where subject groups are preadapted to either high or low offers, we demonstrate that unpredictable shifts in expected offers creates a difference in rejection rates exhibited by the two responder groups for otherwise identical offers. We constructed an ideal observer model that identified neural correlates of norm prediction errors in the ventral striatum and anterior insula, regions that also showed strong responses to variance-prediction errors generated by the same model. Subjective feelings about offers correlated with these norm prediction errors, and the two signals displayed overlapping, but not identical, neural correlates in striatum, insula, and medial orbitofrontal cortex. These results provide evidence for the hypothesis that responses in anterior insula can encode information about social norm violations that correlate with changes in overt behavior (changes in rejection rates). Together, these results demonstrate that the brain regions involved in reward prediction and risk prediction are also recruited in signaling social norm violations.

Introduction

Social norms are standards of behaviors that are based on shared expectations on how individual group members ought to behave in a given situation (Hechter and Opp, 2001). When such expectations are violated, people are willing to forego monetary payoffs to punish the norm transgressors (Fehr and Gächter, 2002). Studies using a simple fairness game, the ultimatum game (Fig. 1*A*), have demonstrated that people reject unfair splits of money even at a cost of themselves, e.g., offers of 20% are rejected about half the time (Camerer, 2003). It has been suggested that the presence of such "irrationality" might be caused by negative emotions such as anger and disgust provoked by unfair treatment (Pillutla and Murnighan, 1995; van't Wout et al., 2006). While these suggestions are reasonable and in many cases compelling, their component parts have not been related to computational models, which could provide new insights into their evolutionary origins and neural implementations.

Montague and Lohrenz (2007) proposed a computational depiction of generating normative behavior and concurrent emotions. To react appropriately in a social exchange, an agent must be able to (1) compute a shared norm about what is expected, (2) detect deviations from that norm, and (3) choose the best actions to correct these deviations. Norms provide baseline (prior) distributions of acceptable signals to be sent to or received from others. In the ultimatum game, the Responder compares the offer observed with the fairness norm and generates error signals carrying information about this deviation from the norm. These error signals include the deviation of the offer from the mean (norm prediction error), and the deviation of the square of the prediction error from the estimated variance (variance prediction error). The norm prediction errors are closely related to the reward prediction errors in gustatory or monetary tasks encoded in the midbrain dopamine neurons (Montague et al., 1996; Schultz et al., 1997; Hollerman and Schultz, 1998; Bayer and Glimcher, 2005; D'Ardenne et al., 2008) and in dopamine-targeted brain areas, such as striatum and orbitofrontal cortex (Pagnoni et al., 2002; McClure et al., 2003; O'Doherty et al., 2003, 2004; Pessiglione et al., 2006). The variance prediction errors resemble the risk prediction errors reported in monetary choice tasks in uncertain environments, involving the anterior insula in particular (Preuschoff et al., 2006, 2008; d'Acremont et al., 2009). When the expectation (norm) is violated, these error signals serve as control signals to guide choices. They may also serve as the progenitor of subjective feelings.

To study the computational substrates of social norm violations, we designed a norm training task using the ultimatum game (Fig. 1) to shift the fairness norm (subject's expectation) so that we were able to quantify norm prediction errors and variance prediction errors. We also recorded the subjective feelings about offers. We hypothesized that the brain areas involved in processing reward prediction errors and risk prediction errors (striatum, anterior insula) are also

Received April 3, 2012; revised Nov. 5, 2012; accepted Nov. 7, 2012.

Author contributions: T.X., T.L., and P.R.M. designed research; T.X., T.L., and P.R.M. performed research; T.X., T.L., and P.R.M. analyzed data; T.X., T.L., and P.R.M. wrote the paper.

This work was supported by National Institute of Mental Health Grant R01 MH085496, National Institute on Drug Abuse Grant R01 DA11723, and the Kane Family Foundation.

*T.X. and T.L. contributed equally to this article.

The authors declare no financial conflict of interest.

Correspondence should be addressed to P. Read Montague, Human Neuroimaging Laboratory, Virginia Tech Carilion Research Institute, 2 Riverside Circle, Roanoke, VA 24016. E-mail: read@vtc.vt.edu.

DOI:10.1523/JNEUROSCI.1642-12.2013

 0270-6474/13/331099-10$15.00/0

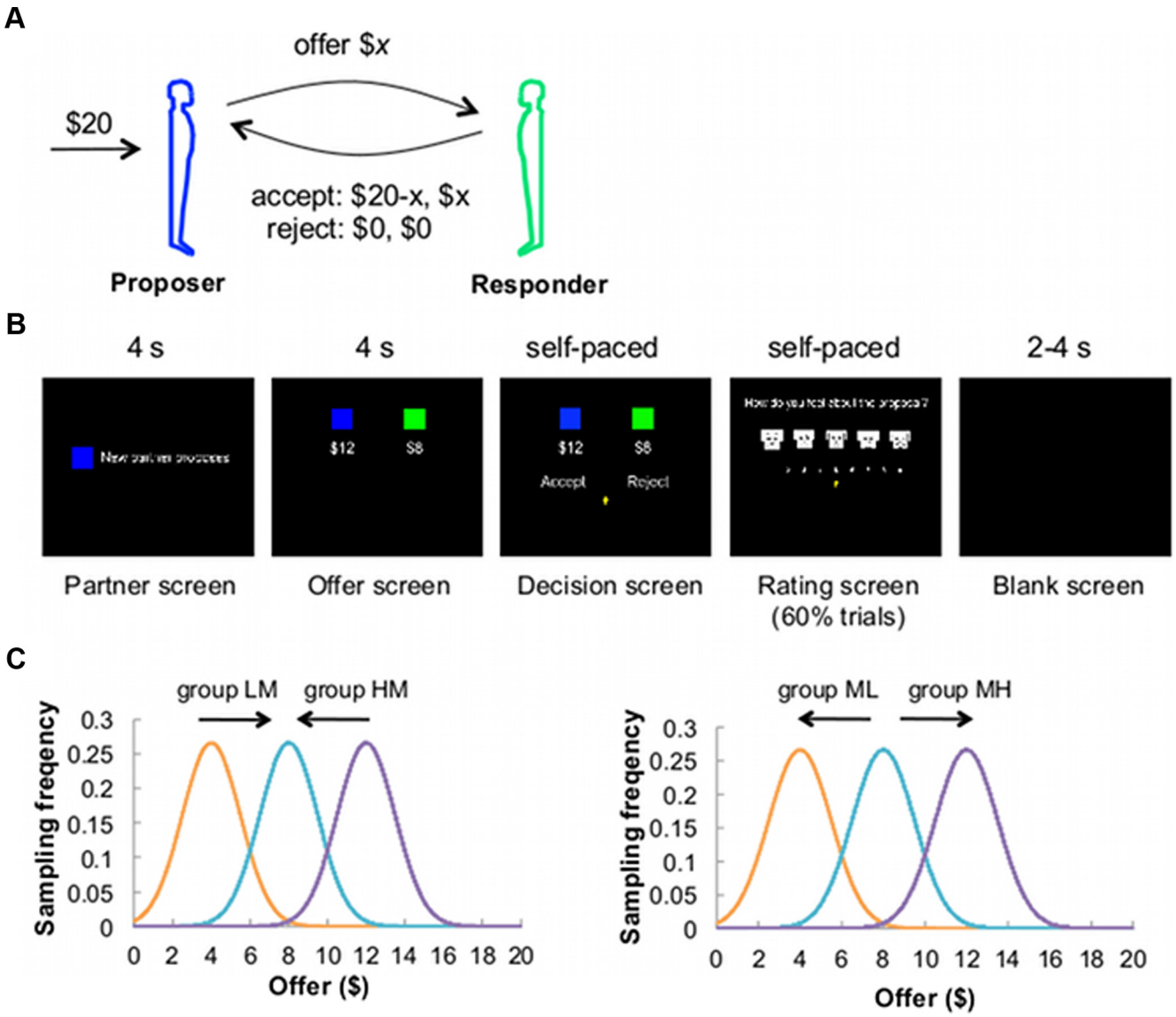

Figure 1. The norm training task. ***A***, The ultimatum game. Subjects played the role of Responder in the ultimatum game. In each round, a new partner (Proposer) made an offer $*x* of $20. Subjects decided to either accept (self got $*x*, partner got $20 − *x*) or reject (both got $0) the split. ***B***, Visual display of the task. Each trial (60 trials in total) began with a new partner (blue square) making an offer (4 s). The offer was displayed for 4 s. Subjects (green square) indicated their decision to accept or reject the offer by moving the yellow arrow (self-paced). On every three of five trials (randomly ordered), subjects were asked to rate their feelings about the offer from 1 (sad face) to 9 (happy face) at a self-paced speed. The intertrial interval was 2–4 s. ***C***, Offers were sampled from one of the three Gaussian distributions, orange curve (mean $4, SD $1.5), cyan curve (mean $8, SD $1.5), and purple curve (mean $12, SD $1.5). Group LM received low offers (orange curve) in the first 30 trials and medium offers (cyan curve) in the last 30 trials. Group HM received 30 high offers (purple curve) first and then 30 medium offers. Groups ML and MH both started with 30 medium offers, but received 30 low offers and high offers, respectively, in the second half of the task.

recruited in signaling norm violations and that reports of subjective feelings are related to these computational signals.

Materials and Methods

Subjects and the norm training task. One hundred twenty-seven subjects (71 females, age: 30.0 ± 8.9 years, age range: 18–59 years) played the ultimatum game while undergoing fMRI scanning. All subjects had normal or corrected-to-normal vision and had no history of neurological or psychiatric disorders. Subjects gave informed consent to participate in the experiments, and all procedures were performed in accordance with the Institutional Review Board of the Baylor College of Medicine.

Subjects played the role of Responder in the game for 60 trials. Each trial started with a new Proposer proposing how to split $20 between the Proposer and the subject (Responder), and ended with the subject's response of accepting or rejecting the offer. If the Responder accepted the offer, both sides got the distributed amounts. However, if the Responder rejected, both sides got $0. To quantitatively manipulate subjects' expectation, we generated the offers from Gaussian distributions with low mean $4, medium mean $8, or high mean $12, and standard deviation $1.5, which was unknown to the subjects. Instead, we instructed the subjects that they were going to play a new, randomly matched partner at each trial. Subjects were paid according to their decisions in two randomly chosen trials and were encouraged to treat each trial as the final outcome. Additionally, at the end of 60% randomly selected trials, we asked the subjects to rate their feelings about the received offers using emoticons ranging from sad to happy on a 1–9 scale. The emoticons were adapted from the self-assessment manikin (Lang, 1980). To ensure sampling emotional ratings at a relatively even pace and without introducing the anticipatory effect, every three of five trials were accompanied with a rating screen. Visual display of the task was back-projected onto a computer screen and viewed through a mirror placed in the scanner. All the choices were made through hand-held button boxes. Stimuli were presented and subjects'

behavioral responses were collected using NEMO (Human Neuroimaging Laboratory, Virginia Tech Carilion Research Institute).

We randomly divided the subjects into four training groups. Group High–Medium (HM; $n = 31$) received high offers in the first 30 trials, sampled from a Gaussian distribution with mean \$12 and standard deviation \$1.5; group Low–Medium (LM; $n = 34$) received low offers in the first 30 trials with mean \$4 and standard deviation \$1.5. Both groups received medium offers in the last 30 trials with mean \$8 and standard deviation \$1.5. Conversely, group Medium–High (MH; $n = 30$) and group Medium–Low (ML; $n = 32$) received 30 medium offers (mean \$8) in the beginning and then 30 high (mean \$12), or low (mean \$4) offers, respectively. Four subjects (2 in group LM, 2 in group HM) were excluded in the imaging analysis due to their excessive movement during scanning. One additional subject in group HM was excluded in the subjective rating analysis because her ratings were not recorded properly.

Bayesian observer model. We modeled subjects throughout the task as Bayesian observers who had a prior of the distribution of offers u, a Gaussian distribution with mean μ, and variance σ^2, denoted as follows: $u \sim N(\mu, \sigma^2)$.

The mean μ and variance σ^2 were also uncertain, and μ and σ^2 were mixed together. Therefore, the prior of offers u is given by:

$$p(u) = p(u \mid \mu, \sigma^2)p(\mu, \sigma^2) = p(u \mid \mu, \sigma^2)p(\mu \mid \sigma^2)p(\sigma^2)$$

When a subject observed a proposal x_t at trial t, he performed the Bayesian update. The posterior was given by:

$$p(u_t \mid x_t) = \frac{p(x_t \mid u_t)\,p(u_{t-1})}{p(x_t)}$$

For convenience, we chose conjugate distributions for μ and σ^2, in which the posterior distribution was in the same family as the prior distribution.

The distribution of mean μ, conditioned on variance σ^2, was given by:

$$\mu \mid \sigma^2 \sim N(\hat{\mu}, \sigma^2/k)$$

The distribution of variance σ^2 took the form of inverse-χ^2 distribution, denoted as:

$$\sigma^2 \sim \mathrm{Inv} - \chi^2(v, \hat{\sigma}^2)$$

The initial values of the hyperparameters μ, k, v, and σ^2 were set as:

$$\hat{\mu}_0 = 10,\ k_0 = 4,\ v_0 = 10,\ \hat{\sigma}_0^2 = 4$$

After observing a proposal x_t at trial t, the values were updated as:

$$k_t = k_{t-1} + 1,\ v_t = v_{t-1} + 1$$

$$\hat{\mu}_t = \frac{k_{t-1}}{k_t}\hat{\mu}_{t-1} + \frac{1}{k_t}x_t$$

$$v_t\hat{\sigma}_t^2 = v_{t-1}\hat{\sigma}_{t-1}^2 + \frac{k_{t-1}}{k_t}(x_t - \hat{\mu}_{t-1})^2$$

Based on the Bayesian observer model, we computed the following parameters used in the imaging analysis:

Expected offer (norm) at trial t: $E[u_t] = \mu_t$

Norm prediction error: $\delta_t = x_t - E[u_{t-1}] = x_t - \mu_{t-1}$

Variance (risk) prediction error (Preuschoff et al., 2008):

$$(x_t - E[u_{t-1}])^2 - E[\sigma_{t-1}^2] = \delta_t^2 - \frac{v_{t-1}}{v_{t-1-2}}\hat{\sigma}_{t-1}^2$$

Positive norm prediction error: $\max(x_t - E[u_{t-1}], 0) = \max(\delta_t, 0)$

Negative norm prediction error: $\max(E[u_{t-1}] - x_t, 0) = \max(-\delta_t, 0)$.

To examine the relationship between the norm prediction errors and the queried subjective feelings, we used the R (R Development Core Team, 2011) function lmer in the R package lme4 (Bates et al., 2011) to perform a mixed linear regression of subjective feelings on the norm and variance prediction error, with random effects (subjects as degree of freedom) on each regressor, including the intercept. Estimation was performed using maximum likelihood. Significance was assessed using the R function ANOVA to compare (likelihood ratio) the full model with the model reduced by each linear regressor in turn (https://stat.ethz.ch/pipermail/r-sig-mixed-models/2009q3/002912.html) (Moore, 2010).

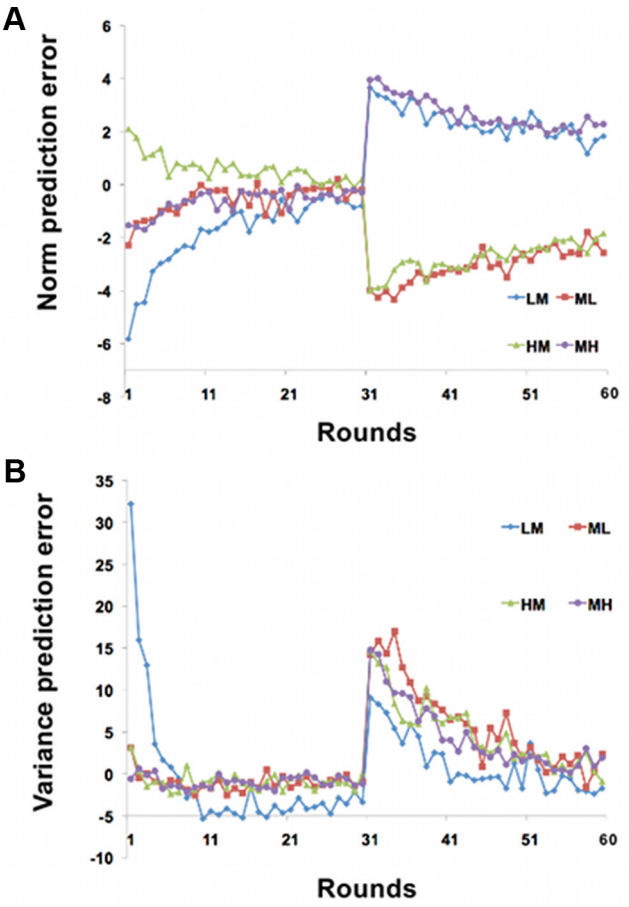

Figure 2. Norm and variance prediction errors. ***A***, Average norm prediction errors by round for each group. ***B***, Average variance prediction errors by round for each group.

To model how subjects made decisions in the task, we combined the inequality aversion model (Fehr and Schmidt, 1999) and the norm-based utility function (Bicchieri, 2006), and defined that the utility of offer x_t was diminished by the degree of norm violations, including both the positive and negative norm prediction errors (after Bayesian update):

$$U(x_t) = x_t - \alpha \cdot \max(E[u_t] - x_t, 0) - \beta \cdot \max(x_t - E[u_t], 0).$$

$\alpha > 0$ is the sensitivity to the negative norm prediction errors (envy); it was discretized in 0.1 increments, ranging from 0 to 10. $0 \le \beta \le 1$ is the sensitivity to the positive norm prediction errors (guilt). It was discretized in 0.1 increments.

We computed the probability of subjects' actions according to the softmax function as follows:

Probability of accepting proposal x_t:

$$P_{\mathrm{accept}} = \frac{e^{U(x_t)/\tau}}{1 + e^{U(x_t)/\tau}}$$

Probability of rejecting proposal x_t:

$$P_{\mathrm{reject}} = \frac{1}{1 + e^{U(x_t)/\tau}}$$

$\tau > (0, 10)$, the temperature, was discretized in 0.1 intervals.

We fitted the above Bayesian observer model to the behavioral data, and estimated the values of α, β, and τ for each subject by maximizing the log likelihood of choices over 60 trials.

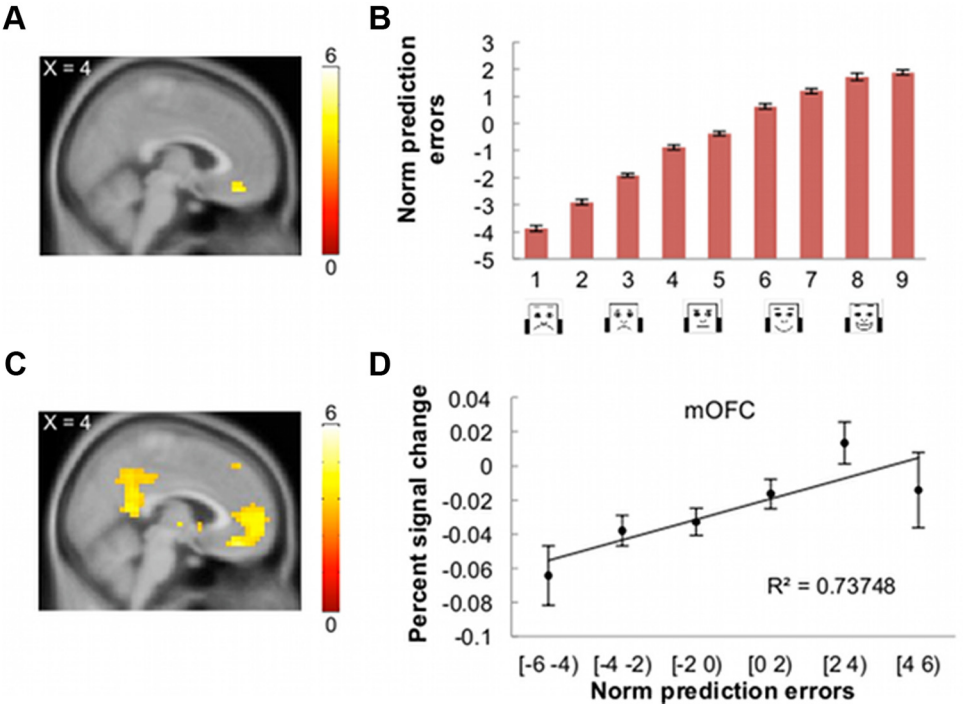

Figure 3. Subjective feelings correlated with norm prediction errors, and both involved the mOFC/vmPFC activation. ***A***, Voxels correlated with norm prediction errors, $p < 0.05$, FDR corrected. Peak voxel (4, 40, −16), $t = 4.38$. ***B***, Emoticon ratings displayed a linear relationship with norm prediction errors. The correlation coefficient was $r = 0.62$. ***C***, Voxels correlated with emoticon ratings of the offers received, $p < 0.01$, FDR corrected. Peak voxel (4, 40, −16), $t = 4.75$. ***D***, ROI analysis using a 6-mm-radius spherical mOFC/vmPFC mask centered on the peak voxel (4, 50, −16) from Harvey et al. (2010). The averaged BOLD response displayed a linear relationship with norm prediction errors. Color bars display *t* scores.

Feeling Bayesian observer. We also used the ideal Bayesian observer model throughout the task to compute the feeling norm and variance prediction errors. The feeling prediction errors were computed in the same fashion as the norm prediction errors, except that the observations were emoticon ratings. Subjects had a feeling prior, a Gaussian distribution with uncertain mean and variance. Subjects performed the Bayesian update upon observing their own emoticon ratings. The updating equations were the same as those for offers. We assumed that the mean of subjects' initial feeling was neutral, taking the value 5. The initial values of the hyperparameters $\hat{\mu}$, k, v, and $\hat{\sigma}^2$ were set as follows: $\hat{\mu}_0 = 5$, $k_0 = 4$, $v_0 = 8$, $\hat{\sigma}_0^2 = 2$.

Image acquisition and analysis. The anatomical and functional imaging was conducted on a 3.0 tesla Siemens Trio scanner. High-resolution T1-weighted scans (1.0 × 1.0 × 1.0 mm) were acquired using an MP-RAGE sequence (Siemens). Functional images were acquired using echo-planar imaging, and angled 30 degrees with respect to the anteroposterior commissural line. The detailed settings for the functional imaging were: repetition time = 2000 ms; echo time = 25 ms; flip angle = 90°; 37 slices; voxel size: 3.4 × 3.4 × 4.0 mm.

Images were analyzed using SPM2 (http://www.fil.ion.ucl.ac.uk/spm/software/spm2/). Slice timing correction was first applied to temporally align all the images. Motion correction to the first functional image was performed using a six-parameter rigid-body transformation. The average of the motion-corrected images was coregistered to each subject's structural images using a 12-parameter affine transformation. Images were subsequently spatially normalized to the Montreal Neurological Institute template by applying a 12-parameter affine transformation, followed by nonlinear warping using standard basis functions. Finally, images were smoothed with an 8 mm isotropic Gaussian kernel and then high-pass filtered (128 s width) in the temporal domain.

General linear models (GLM) were then specified for each subject. All visual stimuli and motor responses were modeled in the design matrix that was constructed by convolving each event onset with a canonical hemodynamic response function in SPM2. Residual effects of head motion were corrected by including the estimated six motion parameters for each subject as covariates. Additional parametric regressors were convolved to the event when offers were displayed and modeled in separate GLM analysis. These regressors of interest included: norm prediction errors (see Fig. 3*A*), emoticon ratings (see Fig. 3*C*), positive norm prediction errors (see Fig. 4*A*), and negative norm prediction errors (see Fig. 4*B*). For the results shown in Figure 5, both norm prediction errors and variance prediction errors were entered in the same GLM at the event of offer revelation without applying orthogonalization. Similarly, for the results shown in Figure 7, both feeling prediction errors and variance prediction errors were entered in the same GLM at the event of offer revelation without applying orthogonalization. Beta maps were estimated for the regressors of interest and then entered into a second-level random effect analysis. Data were displayed using xjView tool box.

Table 1. Estimates from regression of feelings on prediction errors

	Estimate	SE	*t* value	*p* value
Constant	5.352	0.114	47.33	<2e-16
PE	0.616	0.025	24.51	<2e-16
VPE	−0.003	0.006	−0.55	0.585

PE, Prediction error; VPE, variance prediction error. $N = 123$.

In the region of interest (ROI) analysis, 6-mm-radius spherical masks of the medial orbitofrontal cortex (mOFC)/ventromedial prefrontal cortex (vmPFC), anterior insula, and ventral striatum were generated using MarsBaR toolbox (Brett et al., 2002). The signals extracted from the preprocessed images were first averaged within the voxels of the ROI. The spatially averaged signal was linearly detrended for the entire task. Signals time-locked to the event when offers were displayed were generated by linear interpolation. The percentage change in hemodynamic signal was averaged during the 4–8 s period following the offer revelation. The BOLD responses were grouped according to the prediction errors computed from the Bayesian observer model trial by trial at $2 intervals. The mean ± SE of the resulting BOLD signal was plotted in $2 bins for prediction errors by pooling all the trials together.

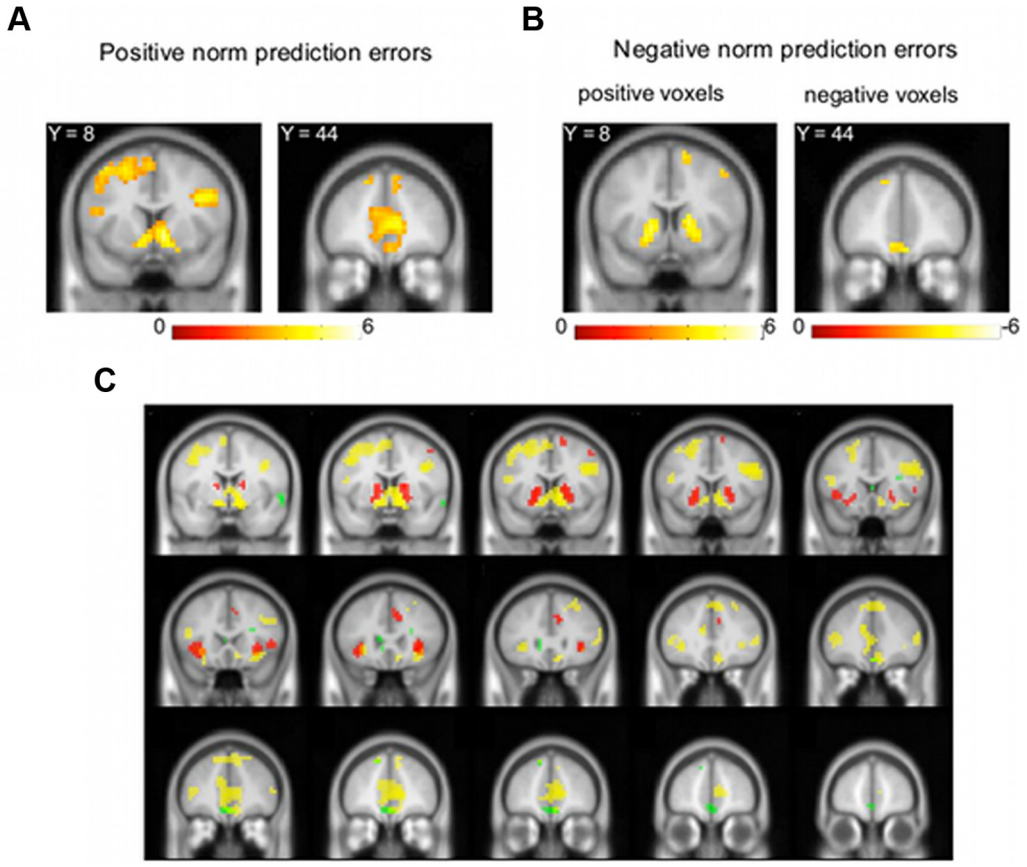

Figure 4. Differential striatal and overlapping mOFC response to positive and negative norm prediction errors. ***A***, Voxels correlated with positive norm prediction errors, $p < 0.05$, FDR corrected. ***B***, Voxels correlated with negative norm prediction errors, $p < 0.05$, FDR corrected. mOFC was negatively correlated with negative norm prediction errors. ***C***, Overlay of voxels from ***A*** (yellow); ***B***, left (red); and ***B***, right (green).

Results

Subjects played the role of Responder in the one-shot ultimatum game (Fig. 1*A*). They were told they would interact with a new partner at each trial for 60 trials in total. The visual stimuli presented at a given trial are shown in Figure 1*B*. We manipulated the fairness norm by randomly assigning subjects into the four groups: LM, ML, HM, and MH (for details, see Materials and Methods). Recall that the LM group saw 30 low offers, followed by 30 medium offers, and similarly for the other three groups. Offers in each condition were sampled from a Gaussian distribution with low ($4), medium ($8) or high ($12) means, as shown in Figure 1*C*.

Norm prediction errors

To quantify the changes in the perception of fairness norm, we modeled an ideal Bayesian observer throughout the task for each subject. For simplicity, we assumed that subjects had the same prior distribution about the fairness norm, a Gaussian distribution with mean $10 and standard deviation $1.5. At each trial, the mean and the variance of the prior distribution were updated according to Bayes rule when a new offer was observed. The posterior was also a Gaussian distribution. The detailed algorithm is included in Materials and Methods. Hence, we had a trial-by-trial measure of the mean and variance for each subject. For convenience, we called the mean of this distribution the norm. The difference between an actual received offer and the expected offer (the prior mean) was the natural prediction error signal, the norm prediction error. Similarly, the variance prediction error was taken as the deviation of the square of the norm prediction error from the estimated variance. Figure 2*A* shows the average prediction error by round, by group, while Figure 2*B* shows the average variance prediction error by round, by group.

We first examined brain regions correlated with norm prediction errors. A random-effect analysis across all subjects ($n = 123$) using norm prediction errors as a regressor at the event of offer presentation throughout the experiment revealed that mOFC/vmPFC activity covaried with norm prediction errors [Fig. 3*A*; $p < 0.05$, whole-brain false discovery rate (FDR) corrected]. As mentioned in the introduction, those norm prediction errors may be related to the subjective feelings about the offers. To examine the relationship between the norm prediction errors and the subjective feelings, we pooled the trials from all the subjects and found the emoticon ratings significantly correlated with norm prediction errors ($r = 0.62$, $p < 10^{-6}$) and they formed a linear relationship (Fig. 3*B*). As the degrees of freedom may be overstated in a pooled regression, we also ran a mixed-effects regression with subject as the degree of freedom and the norm prediction error and variance prediction errors as regressors. The results are summarized in Table 1. The norm prediction error was again highly significant.

We identified brain regions correlated with the subjective feelings by using the emoticon ratings as a regressor at the event of offer presentation in the GLM analysis for each subject throughout the experiment and then performing a random-effect analysis across all subjects ($n = 122$). When the offers were revealed, the activity of vmPFC/anterior cingulate cortex, nucleus accumbens,

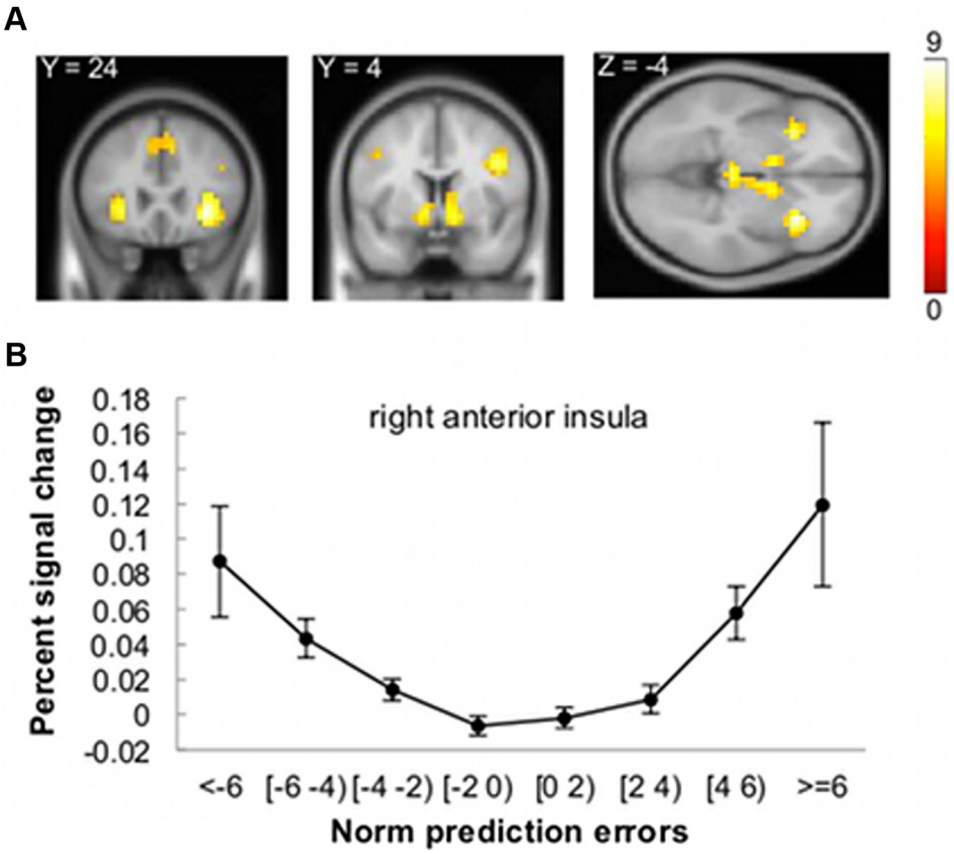

Figure 5. Anterior insula and striatum activity correlated with variance prediction errors. ***A***, Voxels correlated with variance prediction errors, $p < 0.05$, FWE corrected. Right anterior insula, peak voxel (32, 24, −4), $t = 8.70$; right striatum peak voxel (12, 4, −8), $t = 7.10$. ***B***, ROI analysis using a 6-mm-radius spherical mask of the right anterior insula centered on the voxel reported in Preuschoff et al. (2008). The BOLD responses of the right anterior insula displayed a U-shape relationship with norm prediction errors. Color bars display *t* scores.

and posterior cingulate cortex correlated with the emoticon ratings ($p < 0.01$, whole-brain FDR corrected; Fig. 3*C*). The same peak voxel (4, 40, −16) of the mOFC/vmPFC was activated in response to the norm prediction errors ($t = 4.38$; Fig. 3*A*) and the subjective feelings ($t = 4.75$; Fig. 3*C*). To visualize the mOFC/vmPFC activity pattern in terms of norm prediction errors, we performed an ROI analysis using a 6-mm-radius spherical mask centered on the peak voxel of vmPFC reported in another study on preference rating (Harvey et al., 2010), collapsing the trials from all the subjects. The BOLD response of the vmPFC mask displayed a linear relationship with the norm prediction errors (Fig. 3*D*).

We further divided the norm prediction errors into a positive component and a negative component, and ran additional GLM analysis using the positive and negative norm prediction errors as separate regressors. We found that the activity of ventral striatum, vmPFC, and anterior insula correlated with the positive norm prediction errors (Fig. 4*A*; $p < 0.05$, whole-brain FDR corrected). The striatum and anterior insula also correlated with the negative norm prediction errors, but the mOFC negatively correlated with the negative norm prediction errors (Fig. 4*B*; $n = 123$, $p < 0.05$, whole-brain FDR corrected). When overlaying the activity patterns of the positive and negative norm prediction errors, we found that the ventromedial part of striatum specifically correlated with the positive norm prediction errors, whereas a more dorsolateral portion of striatum correlated with the negative norm prediction errors (Fig. 4*C*). But both the anterior insula and the mOFC had overlapping activity patterns. The common region of the mOFC further validated its linear relationship with norm prediction errors, as shown in Figure 3.

We next looked for regions correlated with the variance prediction errors. From a random-effect analysis using variance prediction errors as a regressor, we found that the bilateral anterior insula and ventral striatum had robust responses to risk prediction errors (Fig. 5*A*; $p < 0.05$, whole-brain FWE corrected). We took the mean voxel of the right anterior insula reported to encode risk prediction errors in a financial decision-making task (Preuschoff et al., 2008) and generated a 6-mm-radius spherical mask around it. The ROI analysis showed that the BOLD responses of the anterior insula displayed a U-shape activation pattern to the norm prediction errors (Fig. 4*B*).

Norm training effect

The two groups LM and HM are of particular interest. Group LM ($n = 32$) played 30 offers with low mean \$4 (unfair condition) first, and then switched to play 30 offers with medium mean \$8 (fair condition). Group HM ($n = 31$) played 30 offers with high mean \$12 (hyper-fair condition) and then 30 offers with medium mean \$8 (fair condition). We focused on the latter half of the task when both groups played the fair condition with same offer distributions (Fig. 6*A*). We found significant differences in the rejection rates between the two groups (Fig. 6*B*). Group LM, preconditioned on unfair offers, rejected offers \$6-\$8 less frequently than group HM, preconditioned on hyper-fair offers (Fig. 6*B*). In addition, we asked subjects to rate their feelings about the offers received in 60% of the trials using emoticons ranging from sad to happy. Group LM rated themselves much happier about the medium offers than group HM (Fig. 6*C*).

To examine whether there were any differences in neural responses, we applied the GLM to subjects' fMRI brain images.

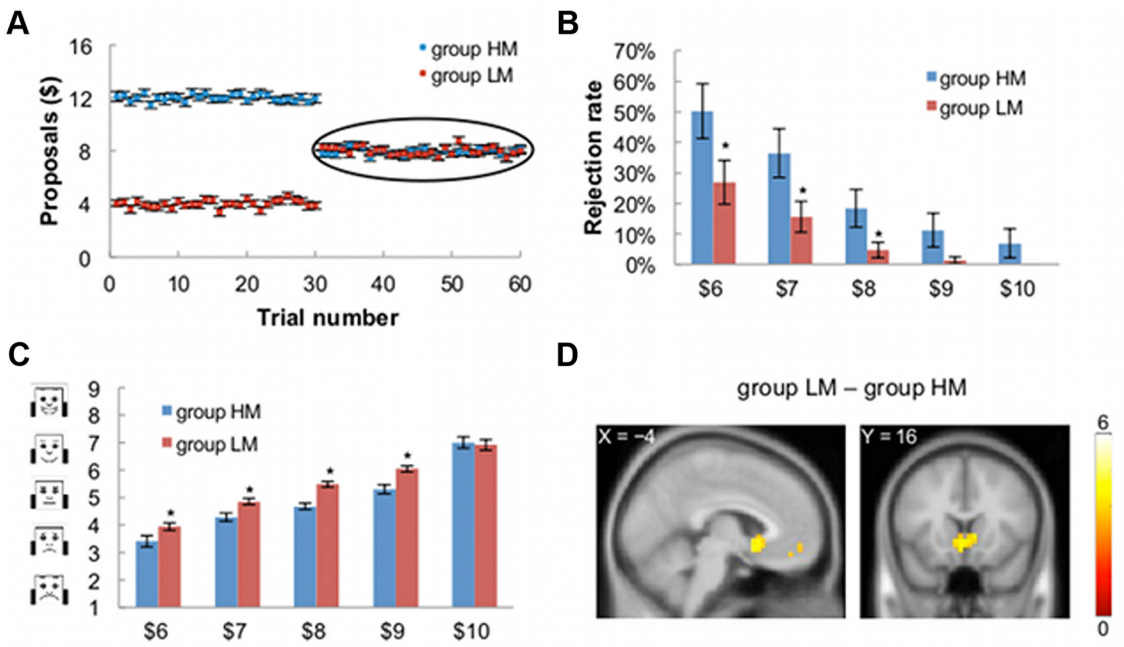

Figure 6. Norm training effect in group HM and LM when both received medium offers. ***A***, Average offers received by the group HM and LM along the course of the task. ***B***, Comparison of the rejection rates between group LM ($n = 34$) and group HM ($n = 31$) when both received medium offers. Group LM preadapted to 30 low offers rejected medium offers \$6–8 less frequently than group HM players who preadapted to 30 high offers, $^*p < 0.05$. ***C***, Comparison of the emoticon ratings between group LM and group HM when both received medium offers. Group LM players rated their feelings about medium offers \$6–9 higher (happier) than group HM players, $^*p < 0.05$. Error bars represent SE. ***D***, SPM contrast between groups LM and HM in the last 30 trials when medium offers were revealed. Group LM had greater activation in the nucleus accumbens and ventromedial prefrontal cortex, $p < 0.001$, uncorrected. Nucleus accumbens (35 voxels), FDR corrected at cluster level, $p < 0.05$. Color bar displays *t* scores.

First, we constructed a contrast of the offer presentation event between group LM and group HM in the fair condition. Neurally, group LM had higher activations in nucleus accumbens and vmPFC than group HM (Fig. 6*D*; $p < 0.001$, uncorrected; nucleus accumbens FDR corrected at cluster level, $p < 0.05$). The activation of the reward regions when observing the offers, the happier feelings toward the offers, and the decreased rejection rates indicate that we were able to change both the perception of the fairness norm (expectation) and the resulting decision-making process.

To relate subjects' choices to the prediction errors, we modeled subjects' choices in the task using a norm-based utility function adapted from Bicchieri (2006) (for details, see Materials and Methods). The commonly used inequity-aversion utility function (Fehr and Schmidt, 1999) is not satisfactory in our situation given that subjects changed their responses to the same unequal outcomes shown in Figure 2. Instead of caring about the unequal divisions between two players, players were sensitive to norm violations, and the utility of an offer received is discounted by its degree of deviations from norm. We considered both the negative and positive deviations from norm in the utility function since hyper-fair offers sometimes were also rejected and subjects reported that they preferred fair offers the most. The envy coefficient, $\alpha \geq 0$, measures a player's sensitivity to negative deviations from norm (negative norm prediction errors). The guilt coefficient, $0 \leq \beta \leq 1$, specifies a player's sensitivity to positive deviations from norm (positive norm prediction errors). We combined the utility function with a logit (softmax) choice function and fitted this model to the actual behavior in the task. For each subject, we estimated his or her envy and guilt coefficients by maximizing the log likelihood of choices over 60 trials. Table 2 presents the summary statistics of the coefficient estimates and log-likelihoods of the fit, and Table 3 gives the average negative log-likelihoods of the fit by group.

Through the norm-based utility function estimated for each subject, we had individual measures of sensitivity to norm violations, i.e., the envy and guilt coefficients. We were interested in finding brain responses modulated by individuals' sensitivity to norm violations. We focused on the negative norm prediction errors and the envy coefficient because only a few subjects had nonzero guilt coefficients. To examine this, we took the beta images corresponding to the negative norm prediction errors and entered them into a second-level analysis using each subject's envy coefficient as a covariate. We found that the dorsal anterior cingulate cortex (dACC) negatively correlated with the sensitivity to negative norm prediction errors (Fig. 7*A*; $p < 0.05$, FDR corrected). We extracted the beta values of the peak voxel (8, 24, 36) from each subject and plotted them against the envy coefficient; the correlation coefficient was $r = -0.36$, $p = 6.24 \times 10^{-6}$ (Fig. 6*B*).

Table 2. Summary statistics of parameter estimates from choice model

	Mean	Median	SD
Envy	3.38	1.75	3.69
Guilt	0.46	0	0.47
Temperature	1.84	0.96	2.15
Log likelihood	8.71	6.01	9.37

$N = 123$.

Table 3. Goodness-of-fit by group for choice model

	LM ($N = 32$)	ML ($N = 32$)	HM ($N = 29$)	MH ($N = 30$)
Average negative log-likelihood	13.57	11.15	5.44	4.12

Total $N = 123$.

Feeling prediction errors

Recent work by Prelec and colleagues (Bodner and Prelec, 2003; Mijović-Prelec and Prelec, 2010) led us to hypothesize that in addition to lower-level mechanisms for tracking parameters of environmentally salient probability distributions, there might be separate

neural mechanisms for monitoring how the subject felt—mechanisms for self-signaling. To probe self-signaling of feelings, we constructed a Bayesian ideal observer model of the subjects' feelings (emoticon ratings) and estimated the feeling prediction errors from the model. The feeling norm prediction error was defined as the difference between the actual reported feeling and the expected feeling. The feeling variance prediction error was the difference between the deviation between the squared feeling norm prediction error and the estimated variance. We found activity in vmPFC, nucleus accumbens, and posterior cingulate cortex correlated with feeling norm prediction errors (Fig. 8*A*; $p < 0.01$, whole-brain FDR corrected). The bilateral anterior insula activity correlated with feeling variance prediction errors (Fig. 8*B*; $p < 0.05$, whole-brain FDR corrected), similar to the norm variance prediction errors. Notably, striatum was not activated by the feeling variance prediction errors.

Discussion

In this work, we used a norm-training task and developed a Bayesian ideal observer model to dynamically track the mean and variance of a norm distribution and serve as a dynamic quantitative probe of norm violations. Using fMRI, we identified brain areas parametrically activated by the learning signals identified by this model. Striatal areas implicated in reward prediction error processing in gustatory or monetary tasks are also activated by norm prediction errors. Similarly, the area activated by risk prediction errors in monetary tasks—bilateral anterior insula (Preuschoff et al., 2008)—is activated by the norm variance prediction error.

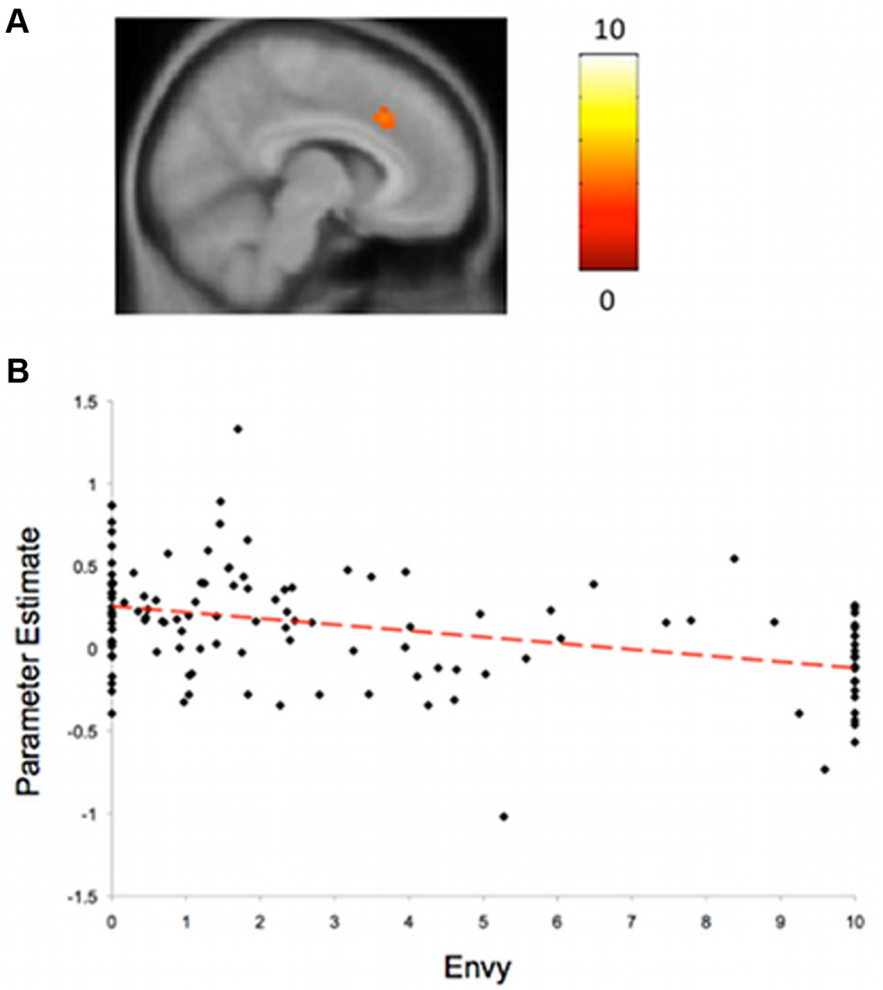

Figure 7. Negative norm prediction error and envy coefficient (alpha). ***A***, Voxels correlated with the individual's sensitivity to the negative norm prediction errors, the envy coefficient (alpha). Second-level analysis on beta images correlated with negative norm prediction errors when offers were revealed. A simple regression using alpha was applied to those beta images. dACC was negatively correlated with alpha, $p < 0.05$, FDR corrected. ***B***, Beta values from the peak voxel (8, 24, 36) of dACC had negative correlation with alpha, $r = -0.39$.

Further, by repeatedly exposing subjects with unfair, fair, or hyper-fair offers, we were able to change subject expectations of offers received in the one-shot ultimatum game without top-down cognitive manipulation. Indeed, subjects' responses to the same offers were modulated by context. We demonstrate that unpredictable shifts in expected offers creates a difference in rejection rates for otherwise identical offers by subject groups pre-adapted to either high or low offers.

We modeled the change in subjects' ongoing expectation of offers using the Bayesian observer model and computed two prediction error signals: the norm prediction error and the variance prediction error. The norm prediction errors predicted subjective feelings as measured by within-task emoticon ratings. Both the norm prediction errors and the subjective feelings recruited the mOFC. The mOFC has been found to represent valuation of both primary and social rewards and mediate hedonic experience (O'Doherty, 2004; Kringelbach, 2005). In addition, mOFC displays a relative coding of value in a context-dependent manner (Seymour and McClure, 2008). Studies have also shown that OFC neurons signal outcome expectancies, which is crucial for adaptive behavior (Schoenbaum et al., 2009). Our finding that the mOFC was activated by both the norm prediction errors and the subjective feelings extends its role to normative decision-making.

The anterior insula and striatum correlated with variance prediction errors, and the anterior insula in particular displayed a U-shape response to the norm prediction errors. It has been shown that anterior insula correlates with the degree of unfairness and predicts the probability of rejecting unfair offers in fairness games (Sanfey et al., 2003; King-Casas et al., 2008). Here we were able to provide a computational account of the role of the anterior insula and extend its role in encoding risk prediction errors during individual financial decision-making to norm violations during social exchange. Together with the norm prediction errors, these results provide evidence that the brain does not just track simple economic variables in the task, but instead builds models of the distribution of rewards and then generates error signals around that distribution. Anterior insula activation has also been associated with interoception and a wide range of emotions including disgust (Damasio et al., 2000; Wicker et al., 2003). It has been recently suggested to play an important role in generating awareness and subjective feelings (Craig, 2002, 2005, 2009). The subjective feelings we queried in the task focused on subjects' preferences about the outcome. Together with its role in risk processing, the anterior insula may respond to emotional

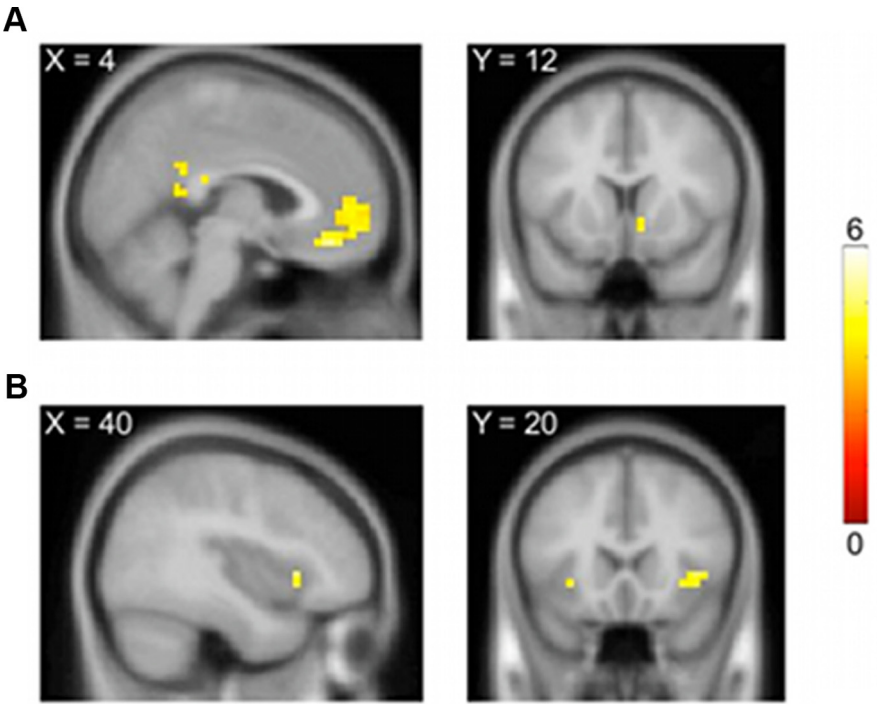

Figure 8. Brain regions correlated with feeling prediction errors. ***A***, Voxels correlated with feeling norm prediction errors, FDR corrected, $p < 0.01$. Nucleus accumbens (8, 12, −8), $t = 3.96$; vmPFC (4, 40, −16), $t = 5.01$; posterior cingulate (−4, −32, 40), $t = 4.65$. ***B***, Voxels correlated with feeling variance prediction errors, FDR corrected, $p < 0.05$. Right anterior insula (40, 20, 0), $t = 4.69$; left anterior insula (−32, 28, 4), $t = 4.30$. Color bars display t scores.

arousal evoked by deviations from norm, which we did not assess in this study. Patients with insular lesions have reported reduced arousal in response to both unpleasant and pleasant pictures (Berntson et al., 2011).

We also observed a differential activation pattern in striatum to the positive and negative norm prediction errors. The ventromedial part of striatum specifically correlated with the positive norm prediction errors, whereas a more dorsolateral portion of striatum correlated with the negative norm prediction errors. The ventromedial part of striatum receives inputs from vmPFC and OFC, and the more dorsolateral part of striatum receives inputs from dACC (Voorn et al., 2004; Haber and Knutson, 2010). This anatomical characterization matches the functional activity patterns identified from the GLM analysis. Indeed, the vmPFC/OFC coactivated with the ventromedial striatum in response to the positive norm prediction errors, whereas the dACC coactivated with the more dorsolateral region of striatum in response to the negative norm prediction errors. The role of dACC involves conflict monitoring and cognitive control (Botvinick et al., 1999, 2004; Kerns et al., 2004). We also found that dACC activity tended to have higher response to negative norm prediction errors in subjects with a low-envy coefficient (rational, accepted more unfair offers). This suggests that those subjects experienced bigger conflicts and engaged more cognitive control when receiving unfair offers than subjects with high-envy coefficients (irrational, rejected unfair offers frequently).

Like others (Fehr and Gächter, 2002; Sanfey et al., 2003; Boyd et al., 2010), we suggest that the rejection rates and changes in rejection rates observed during our experiment provide strong evidence that this paradigm probes social-processing mechanisms. For example, subjects reject positive-valued offers throughout the entire experiment—both during the preconditioning (e.g., on the high offer distribution) and during postconditioning (e.g., on the medium offer distribution). In the case of a rejected offer, the subject receives no reward. There is no reward-harvesting rationale for a human to reject nonzero offers; therefore, we find it problematic to argue that we are observing strictly reward-related responses when no reward is received and this nonreceipt is due to a choice made by the subject. If subjects were passive recipients of the proposals, then one could argue that they might simply be building models of the proposal distributions and their brains are responding to fluctuations from expectations of the model. However, in the current design, the subjects willfully choose whether to accept a proposal so that a receipt of 0 reward is due to the subject, not the proposal size. This too would not be the behavior of a simple reward-gathering agent.

Instead, we suggest that the positive rejection rates seen throughout the experiment and the large change in rejection rate when the distributions are switched are best explained by a social mechanism that seeks to send a signal back to the proposer. The well reported need to punish unfairness in other humans (Fehr and Gächter, 2002; Boyd et al., 2010) also supports a social mechanism in this experiment. This discussion raises the issue of how an optimal reinforcement-learning agent might execute this task.

Absent some assumption about future consequences of accepting offers, any reasonable reward-harvesting reinforcement learning model would, either through prior assumptions or through learning, accept all offers. There is no reason short of some kind of further model of the other player that a shift in the offer distributions should cause a rejection rate change. Even if a reinforcement learning model was building a sufficient statistic model of the offer distributions, there is no incentive to ever reject an offer.

A prominent theory of decision making in fairness games rests on the concept of inequity aversion (Fehr and Schmidt, 1999). In general, inequity-averse players care about the differences in payoffs between self and partner, which motivates the behavior observed, including rejection in the ultimatum game. Fehr and Schmidt (1999) commit to a specific form of a utility function using an absolute level of inequality that captures these ideas. Our results concerning the differing rejection rates in the separately trained groups after the offer distribution switch suggest that this inequity aversion theory is inadequate. Chang and Sanfey (2011) have recently reported a similar finding, also using a model of behavior incorporating deviations from expectation. Importantly, however, these investigators established the subjects' expectations beforehand and did not dynamically model changes in expectations. While Chang and Sanfey (2011) and our results both suggest that the Fehr–Schmidt model is inadequate, our result that rejection rates for offers from a fair distribution are different for differentially conditioned groups shows that a dynamic, learning approach is needed for understanding social norms.

Finally, also using a Bayesian ideal observer model of the subjects' feelings, we explored the possibility that there might be separate neural mechanisms for monitoring how the subject felt—mechanisms for self-signaling (Bodner and Prelec, 2003; Mijović-Prelec and Prelec, 2010), in addition to lower-level

mechanisms for tracking parameters of environmentally salient probability distributions. We found that the feeling norm prediction error tracked similar regions in vmPFC and striatum as the norm prediction error, but that the feeling variance prediction error recruited anterior insula, but not striatum, suggesting a physiological dissociation between these types of error signals.

Social dysfunction is a defining feature of psychiatric disease and involves both cognitive and emotional impairments. As such, norm processing may come to play an important role in understanding psychiatric disease. A previous study has shown that norm processing is impaired in individuals with borderline personality disorder (King-Casas et al., 2008). In this paper, we illuminate the neural substrates of a computational depiction of social norm violations. These neural signals may find use as objective biomarkers helping characterize mental disorders (Kishida et al., 2010; Montague et al., 2012).

References

Bates D, Maechler M, Bolker B (2011) lme4: linear mixed-effects models using S4 classes. R package version 0.999375-40. http://CRAN.R-project.org/package=lme4.

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141. CrossRef Medline

Berntson GG, Norman GJ, Bechara A, Bruss J, Tranel D, Cacioppo JT (2011) The insula and evaluative processes. Psychol Sci 22:80–86. CrossRef Medline

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

Bodner R, Prelec D (2003) Self-signaling in a neo-Calvinist model of everyday decision making. In: Psychology of economic decisions, vol. 1 (Brocas I, Carillo J, eds), pp 105–126. London UK: Oxford UP.

Botvinick M, Nystrom LE, Fissell K, Carter CS, Cohen JD (1999) Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature 402:179–181. CrossRef Medline

Botvinick MM, Cohen JD, Carter CS (2004) Conflict monitoring and anterior cingulated cortex: an update. Trends Cogn Sci 8:539–546. CrossRef Medline

Boyd R, Gintis H, Bowles S (2010) Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science 328:617–620. CrossRef Medline

Brett M, Anton JC, Valabregue R, Poline JB (2002) Region of interest analysis using an SPM toolbox. Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, Sendai, Japan. Available on CD-ROM in Neuroimage 16:2.

Camerer CF (2003) Behavioral game theory: experiments in strategic interaction. Princeton, NJ: Princeton UP.

Chang LJ, Sanfey AG (2011) Great expectations: neural computations underlying the use of social norms in decision-making. Soc Cogn Affect Neurosci. Advance online publication. doi:10.1093/scan/nsr094. CrossRef Medline

Craig AD (2002) How do you feel? Interoception: the sense of the physiological condition of the body. Nat Rev Neurosci 3:655–666. Medline

Craig AD (2005) Forebrain emotional asymmetry: a neuroanatomical basis? Trends Cogn Sci 9:566–571. CrossRef Medline

Craig AD (2009) How do you feel–now? The anterior insula and human awareness. Nat Rev Neurosci 10:59–70. CrossRef Medline

d'Acremont M, Lu ZL, Li X, Van der Linden M, Bechara A (2009) Neural correlates of risk prediction error during reinforcement learning in humans. Neuroimage 47:1929–1939. CrossRef Medline

Damasio AR, Grabowski TJ, Bechara A, Damasio H, Ponto LL, Parvizi J, Hichwa RD (2000) Subcortical and cortical brain activity during the feeling of self-generated emotions. Nat Neurosci 3:1049–1056. CrossRef Medline

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science 319:1264–1267. CrossRef Medline

Fehr E, Gächter S (2002) Altruistic punishment in humans. Nature 415:137–140. CrossRef Medline

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

Haber SN, Knutson B (2010) The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology 35:4–26. CrossRef Medline

Harvey AH, Kirk U, Denfield GH, Montague PR (2010) Monetary favors and their influence on neural responses and revealed preference. J Neurosci 30:9597–9602. CrossRef Medline

Hechter M, Opp KD (2001) Social norms. New York: Russell Sage Foundation.

Hollerman JR, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. Nat Neurosci 1:304–309. CrossRef Medline

Kerns JG, Cohen JD, MacDonald AW 3rd, Cho RY, Stenger VA, Carter CS (2004) Anterior cingulated conflict monitoring and adjustments in control. Science 303:1023–1026. CrossRef Medline

King-Casas B, Sharp C, Lomax-Bream L, Lohrenz T, Fonagy P, Montague PR (2008) The rupture and repair of cooperation in borderline personality disorder. Science 321:806–810. CrossRef Medline

Kishida KT, King-Casas B, Montague PR (2010) Neuroeconomic approaches to mental disorders. Neuron 67:543–554. CrossRef Medline

Kringelbach ML (2005) The human orbitofrontal cortex: linking reward to hedonic experience. Nat Rev Neurosci 6:691–702. CrossRef Medline

Lang PJ (1980) Behavioral treatment and bio-behavioral assessment: computer applications. In: Technology in mental health care delivery systems (Sidowski JB, Johnson JH, Williams TA, eds), pp 119–137. Norwood, NJ: Ablex.

McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. Neuron 38:339–346. CrossRef Medline

Mijović-Prelec D, Prelec D (2010) Self-deception as self-signalling: a model and experimental evidence. Philos Trans R Soc Lond B Biol Sci 365:227–240. CrossRef Medline

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

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947. Medline

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

Moore C (2010) Linear mixed-effects regression *p*-values in R: a likelihood ratio test function. http://blog.lib.umn.edu/moor0554/canoemoore/2010/09/lmer_p-values_lrt.html

O'Doherty JP (2004) Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr Opin Neurobiol 14:769–776. 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

O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454. CrossRef Medline

Pagnoni G, Zink CF, Montague PR, Berns GS (2002) Activity in human ventral striatum locked to errors of reward prediction. Nat Neurosci 5:97–98. CrossRef Medline

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behavior in humans. Nature 442:1042–1045. CrossRef Medline

Preuschoff K, Bossaerts P, Quartz SR (2006) Neural differentiation of expected reward and risk in human subcortical structures. Neuron 51:381–390. 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

Pillutla MM, Murnighan JK (1995) Being fair or appearing fair: strategic behavior in ultimatum bargaining. Acad Manage J 38:1408–1426.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the ultimatum game. Science 300:1755–1758. CrossRef Medline

Schoenbaum G, Roesch MR, Stalnaker TA, Takahashi YK (2009) A new

perspective on the role of the orbitofrontal cortex in adaptive behavior. Nat Rev Neurosci 10:885–892. CrossRef Medline

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599. CrossRef Medline

Seymour B, McClure SM (2008) Anchors, scales and the relative coding of value in the brain. Curr Opin Neurobiol 18:173–178. CrossRef Medline

van't Wout M, Kahn RS, Sanfey AG, Aleman A (2006) Affective state and decision-making in the Ultimatum Game. Exp Brain Res 169:564–568. CrossRef Medline

Voorn P, Vanderschuren LJ, Groenewegen HJ, Robbins TW, Pennartz CM (2004) Putting a spin on the dorsal-ventral divide of the striatum. Trends Neurosci 27:468–474. CrossRef Medline

Wicker B, Keysers C, Plailly J, Royet JP, Gallese V, Rizzolatti G (2003) Both of us disgusted in my insula: the common neural basis of seeing and feeling disgust. Neuron 40:655–664. CrossRef Medline