Neural Responses to Sanction Threats in Two-Party Economic Exchange

Jian Li, Erte Xiao, Daniel Houser and P Read Montague

June 2009

Discussion Paper

Interdisciplinary Center for Economic Science
4400 University Drive, MSN 1B2, Fairfax, VA 22030
Tel: +1-703-993-4850 Fax: +1-703-993-4851
ICES Website: www.ices-gmu.org
ICES RePEc Archive Online at: http://edirc.repec.org/data/icgmuus.html

Neural Responses to Sanction Threats in Two-Party Economic Exchange

Jian Li[*], Erte Xiao[‡], Daniel Houser[§] & P Read Montague[*¶†]

June, 2009

[*]Department of Psychology, New York University, New York, NY 10003, USA. [‡]Department of Social and Decision Sciences, Carnegie Mellon University, USA. [§]Interdisciplinary Center for Economic Science (ICES), George Mason University, Fairfax, VA 22030, USA. [¶]Menninger Department of Psychiatry & Behavioral Sciences, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.

[†]Correspondence should be addressed to P.R.M. (rmontague@hnl.bcm.edu) Menninger Department of Psychiatry & Behavioral Sciences, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.

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

Acknowledgements

This research was supported by the Kane Family Foundation (P.R.M.), NINDS grant NS-045790 (P.R.M.) and NIDA grant DA-11723 (P.R.M.). We thank N. Apple for experimental design, S.M. McClure for helpful discussions, C. Bracero & J. McGee for fMRI images collection. Detailed comments from two anonymous referees substantially improved this paper.

Competing interests statement

The authors declare no competing interests.

Abstract

Sanctions are used ubiquitously to enforce obedience to social norms. Recent field studies and laboratory experiments have demonstrated, however, that cooperation is sometimes reduced when incentives meant to promote pro-social decisions are added to the environment. Although a variety of explanations have been suggested, the neural foundations of this effect have not been fully explored. Using a modified trust game, we find trustees reciprocate relatively less when facing sanction threats, and the presence of sanctions significantly reduces trustee's brain activities involved in social reward valuation (VMPFC, LOFC, and Amygdala), while simultaneously increases brain activities in parietal cortex previously implicated in rational decision making. Moreover, we find that neural activity in trustee's VMPFC area predicts her future level of cooperation under both sanction and no-sanction conditions, and that this predictive activity can be dynamically modulated by the presence of a sanction threat.

Introduction

Sanctions are ubiquitous in modern human societies[1]. The purpose of sanctions is to enforce norm-obedience beyond the level that humans might achieve in the absence of punishment[2-4]. Several recent field studies and laboratory experiments have established, however, that adding monetary sanctions to an environment can reduce cooperation[5-7]. Substantial speculation has arisen surrounding the source of this counter-intuitive effect, including that the presence of sanctions might change individuals' perceptions of the environment, thus crowding out internal motivations for cooperation (e.g., a preference to obey social norms[5-8]). Imposing sanctions also can be seen as a signal of distrust[9-11], or might create a hostile atmosphere[12-13] and for these reasons reduce cooperation.

Previous behavioral experiments have sought to distinguish these competing explanations. For example, a recent study[5] reported data from an experiment aimed at determining the relative importance of intentions and incentives in producing non-cooperative behavior. Participants played a one-shot investment experiment in pairs. Investors sent a certain amount to trustees, requested a return on that investment and, in some treatments, could threaten sanctions to enforce their requests. Decisions by trustees facing threats imposed (or not) by investors were compared to decisions by trustees facing threats imposed (or not) by nature. The main finding was that, when not threatened, trustees typically decided to return a positive amount less than the investor requested. When threatened, however, that decision became least common. Moreover, those findings do not vary with whether the sanction was intentionally imposed by the investor or by nature. The results are consistent with the view that the detrimental effect of sanctions on cooperation might not hinge on trustees' perceptions of investor intentions. Here we provide novel neurological evidence that contributes to an improved understanding of the biological mechanisms underlying sanction effects on pro-social decisions.

We report results from a functional magnetic resonance imaging (fMRI) study with an investment game that offers new data on the source of detrimental sanction effects. We examine the specific hypothesis that sanctions change individuals' perceptions of the environment[5-8]. The perception shift hypothesis has been elucidated in detail by others[5]. In brief, absent external incentives, people are hypothesized to justify their behavior through an appeal to internal social motivations. However, when a threat of sanctions is present, cognitive dissonance theory suggests this external incentive can become a salient behavioral justification.

In particular, due to self-serving bias, it's likely for people to interpret the fine as a price in order to support selfish decisions based on an analysis of personal costs and benefits. Thus, neural support for this hypothesis is found if sanction threats encourage activity within neural networks associated with self-interested economic decision making[14-16], while simultaneously mitigating activity in brain networks implicated in social reward evaluation and processing[17-26]. Those latter networks, on the other hand, are hypothesized to be relatively more active when sanctions are not used.

The specific areas of the brain of interest to the "perception shift" hypothesis are reasonably well established. Engagement of parietal cortex in self-interested economic decision-making, and especially expected utility calculations, has been indicated using various experimental paradigms in both primate and human neuroimaging studies[14-16]. Neural networks involved in social rewards have also been heavily researched[17-29]. Of particular interest to us is the orbitofrontal cortex (OFC), since it is well known to be reliably involved in processes including social reward evaluation and decision-making[15, 17, 19, 20, 29-32]. However, despite the substantial literature in neuropsychology and psychiatry pointing to the importance of prefrontal cortex and the OFC in social recognition and interaction[19, 21-25, 33, 34], ours are among the first experiments informing OFC's role in perceiving and evaluating threats of sanctions. In particular, we investigate (i) how activation patterns in OFC depend on whether one is threatened with sanctions and (ii) to shed light on whether the activity of the medial area of OFC, the ventromedial prefrontal cortex (VMPFC), a brain area that seems pivotal in human decision-making[15, 17, 18, 35-39], also predicts subjects' social exchange decisions.

Improving our understanding of the role of OFC is critical in furthering our knowledge of human reward valuation and social decision-making. It is known that humans with impairments to this region exhibit emotional and social behavioral dysfunctions characterized by impulsive behavior, poor insight, and inappropriate social and financial decision-making[31, 35-37]. Also, patients with OFC damage have difficulty recognizing and acting upon stimulus-reward/punishment contingencies and adaptively adjusting their behaviors to maximize monetary rewards[35]. Data also indicate subjects with OFC lesions do not change their decisions in response to variations in levels of ambiguity and risk[31]. Furthermore, emerging evidence from single-electrode neurophysiology recordings in primates, as well as human neuroimaging studies, indicate that neural activities in the OFC could be well-suited for representation of primary and abstract reinforcers[31, 32, 35, 36] (including reward and punishment.)

Our study used the event-related fMRI technique together with an investment game. The game has previously been reliably used to elicit detrimental sanction effects[5,9] (Fig. 1*A* and 1*B*, also see supporting information (SI) Fig. 1). Two mutually anonymous participants were paired together for 10 trials (both players were given 10 monetary units (MU) at the beginning of each trial) and were assigned to the role of investor and trustee, respectively (supporting information (SI) Fig.1 and 2). Subject pairs, as well as subjects' roles within each pair, remain fixed for the entire 10 rounds. The investor moves first and makes three consecutive decisions. First the investor decides how many monetary units to send to the trustee. The investor knows that any amount sent is tripled by the experimenter and then given to the trustee. Next the investor decides on the number of monetary units to request back from the trustee. Finally, the investor decides whether to threaten (credibly) a sanction. If the investor chooses the sanction option and the trustee fails to return the requested amount to the investor then a fixed amount of monetary units (4 MU) are deducted from the trustee's final earnings. The investor's three decisions are revealed to the trustee, and then the trustee decides on the amount to return to the investor (see also supporting information (SI) Fig. 1).

Although our game was repeated 10 times, we first derive standard Nash equilibrium predictions based on selfish preferences for the one-shot game. In this environment, trustees should not return any amount if the investor does not impose a sanction threat. Consequently, the investor should send nothing, meaning that both would earn their endowment of 10 MU. However, threatening a sanction of 4 MU can enforce a backtransfer request of at most four. Thus, a Nash equilibrium in this case occurs when an investor sends 1 (or 2) MU, requests a backtransfer of 3 (or 4) MU and threatens a sanction of four. The trustee then returns 3 (or 4) MU to the investor. In both cases, the investor earns 12 MU. The trustee earns 10MU when the investor sends one and 12MU when the investor sends two. Thus, there are multiple Nash equilibrium for the one-shot game, and trustees are predicted to return more under a punishment threat (return three or four) than when punishment is not threatened (return zero).

One Nash equilibrium (NE) for the repeated game involves playing any one-shot equilibrium in every round. Previous studies, however, find investors send more, and trustees subsequently return more, than would be predicted by such a "naïve" equilibrium[5,9]. One possible explanation for this is that fully-rational participants play a "sophisticated" subgame perfect NE strategy that leads to greater amounts sent and returned than predicted by one-shot equilibria. Such "as if"

cooperative equilibria can exist when the one-shot game admits two or more Nash equilibria, and when one of these equilibria is "worse" for the players than another[40,41].

In our game the NE payoff of (12,10) is worse for a trustee than (12,12), and this can lead to "as if" cooperative behavior. To see how, suppose that instead of ten times our game was repeated only twice. Then, for example, in the first round the investor could send three, ask for a return of six, and threaten to punish. If the trustee reciprocates and returns the full six (implying that both earn 13 in the first round) then the investor plays the "nice" NE in the final round, meaning the trustee earns 12 for a total payoff of 25. If the trustee instead defects in the first round, then the investor plays the (12,10) equilibrium in the final round. It follows that the trustee has no incentive to defect in the first round, and "as if" cooperative equilibrium behavior obtains. By similar reasoning, one can see that "as if" cooperation can be supported as a sophisticated NE of our ten round game. We demonstrate below, however, that the cooperative patterns found in our data are inconsistent with this explanation (see **Results**).

Our goal with this research is to shed light on why investors and trustees send more than predicted by standard economic theory, and to help explain why trustees return relatively less to investors when threatened by sanctions than when not[5,9]. Our hypothesis is that the detrimental effect of sanctions is due to a "perception shift" where the sanction becomes the "price" for selfishness, changing the decision context from a non-market social environment where giving activates social reward networks, to a market-based exchange centered on maximizing personal benefits[5]. As noted above, this hypothesis has clear neural implications. In particular, ventromedial prefrontal cortex (VMPFC), lateral orbitofrontal cortex (LOFC) and amygdala, all of which are reliably involved in processing abstract and social reward processing in human brain, are expected to be relatively more strongly activated in the social-environment created by the absence of sanctions. Under sanction threats, and thus a market-decision context, neural networks that encode individual expected utility (e.g., bi-lateral parietal cortex) should be relatively more active.

We collected continuous blood-oxygen-level dependent (BOLD) images from trustees while they made decisions in the investment game. Investor brain activity was not monitored. Because participants played the game in fixed pairs, reputation could presumably accumulate throughout the experiment. This presents no difficulties for our analysis because we focus on sanction vs. no-sanction contrasts across all rounds and subjects, thus controlling any reputation effects. In

particular, as described below, our design enables us to separate effectively trustees' neural responses to each of the three decisions made by investors, and thus allows us to distinguish the separate correlates of these neural patterns with trustees' decisions. Our analysis finds support for the perception shift hypothesis, and suggests further that VMPFC may integrate incentives represented by activations of multiple neural networks.

Results

Investigating whether subjects use sophisticated Nash equilibrium strategies

We describe below that investors send, and trustees return, substantial amounts. To shed light on whether this seemingly cooperative behavior might stem from sophisticated non-cooperative NE strategies, observe that a one-shot NE *must* be played in the last period of any sophisticated equilibrium path[41]. The average amount invested in the final round of our game is 5.9. This amount is statistically significantly larger than two ($p<0.001$, two-sided t-test), which is the largest possible equilibrium investment amount in the one-shot game. Indeed, the vast majority (over 2/3) of investors send more than two in the final round and, of those who send two or less, 41% make a punishment or backtransfer request decision that is inconsistent with a one-shot equilibrium. Thus, only about 15% of investors in the final round of our game make decisions consistent with sophisticated NE. Moreover, the average amount returned by trustees in the final round is 9.7. This amount is statistically significantly larger than four ($p<0.001$, two-sided t-test), which is the maximum return consistent with equilibrium in the stage game.

Finally, note also that sophisticated equilibria can involve "trigger strategies," under which an investor reverts to a one-shot NE following a defection. In our data, however, average investment following a defection (returning less than the investor requested) is 5.4, which is more than half of the endowment and again is statistically significantly larger than two ($p < 0.001$, two-sided t-test.)

In light of our evidence, we conclude that sophisticated NE play is not a plausible explanation for the cooperative patterns found in our data.

Sanction decisions and their effect on trustees' repayment decisions

On average, investors imposed threats of sanctions 49.3% of the time following a trustee's decision to defect, while the frequency following cooperation was similar at 46.0%. Out of 52 investors, eight imposed sanctions on every trial, while 11 never imposed a sanction. Overall, an investor's decision to impose a threat was uncorrelated with whether a trustee defected in the previous period (two-sample χ^2 test, p= 0.78). However, investors were more likely to use sanctions in a given trial if (i) his trustee defected in the immediately previous trial and (ii) a sanction had not been used in that previous trial (χ^2 = 23.38, p = 0.001). Overall, investors chose the sanction option 46.3% of the time, ranging from a high of 53.7% (round 9) to a low of 37.0% (round 1). Using a mixed effect analysis including a one sample t-test and a logistic regression, we found the correlation between the use of sanctions and the round did not survive statistical thresholds (average sigmoid slope = 1.64, p=0.053). Three important variables: investor's investment (mean slope = -0.048, p = 0.52), investor's request (mean slope = -0.013, p = 0.87) and trustee's repayment (mean slope = -0.03, p = 0.64) are not correlated with round number.

To assess trustees' behavioral responses to sanction threats, we first plot as a baseline an "equal split" strategy (Fig. 2*B*, dotted line). This strategy could emerge if a trustee treats the tripled investment amount as a common good and demands half of it. We compare this to trustees' mean real repayments under sanction threats (Fig. 2*B*, blue line) and when not threatened with sanctions (Fig. 2*B*, red line). The vertical lines in the figures are one s.e. of the trustees' mean repayment in both conditions. The trustee's repayment when threatened with sanctions is significantly different between cases where sanctions are and are not imposed (two-sample t-test, p<0.05, also see supporting information (SI) Fig. 3 and SI Table 1 for details). The difference is greater when the investments are larger (greater than six). Overall, trustees' average repayments in sanction and no-sanction cases are 6.05 monetary units (MU) and 12.04 MU, respectively (SI Table 1). Thus, the difference in repayment amounts cannot be explained only by the possibility that trustees choose to keep 4 MU extra in the sanction condition as compensation for the sanction's cost.

Previous research suggests that trustees' repayments might also depend on whether the investor used the sanction to enforce an "unfair" backtransfer request[5] (defined as a request for greater than 2/3 of the tripled investment amount, which is the amount that equalizes investor and trustee earnings.) To investigate unfair requests, we first explored investor behavior by

plotting the backtransfer request against the investment decision for both the sanction and no-sanction conditions (Fig. 2*A*, blue and red lines). The dotted line in that figure indicates a request of 2/3 of the tripled investment. It is apparent that the investors' requests do not differ significantly between the sanction and no-sanction conditions (t-test, p=0.9), nor are the averages significantly different on average from equal-earnings requests ($p_{\text{no-sanction}} = 0.9$, $p_{\text{sanction}} = 0.9$).

With respect to trustees' decisions, consistent with previous studies[5] we find sanctions to have a detrimental effect on trustees' returns both when the investor's back-transfer request is fair as well as when it is unfair, and we find that these detrimental effects are not statistically significantly different. In particular, a fair request results in a mean return equal to 53% of the tripled investment amount, while combining sanctions with a fair request reduces returns to 47% on average. When the request is unfair the analogous change is from 59% to 47%, and this between-condition difference (a six vs. 12 percentage point reduction) is not statistically significant (Wilcoxon test, p>0.15, two-tailed).

Trustees' neural responses to the revelation of sanctions

The previous sections detailed the effect of sanctions on trustees' repayment decisions. To shed light on the neural underpinnings of this effect we used a standard general linear model analysis (GLM) to compare trustees' brain responses between cases where sanctions were and were not threatened by the investor. The sanction – no-sanction contrast did not identify any prefrontal brain activities at p<0.001 level (uncorrected; 5 continuous voxels; see SI Table 3). However, the no-sanction – sanction contrast revealed differential activation in areas implicated in social reward processing (Fig. 3, SI Table 2). These brain areas (Fig 3, p<0.001; 5 continuous voxels; uncorrected) include the ventromedial prefrontal cortex (VMPFC, peak activity at MNI [4 56 -4]), the lateral orbitofrontal cortex (LOPFC, peak activity at MNI [32 52 -4]), the posterior cingulated cortex (PCC, peak activity at MNI [4 -24 36]), and the right amygdala (peak activity at MNI [24 0 -20]). We conducted a region of interest (ROI) analysis to further investigate these results (Fig. 3*B*). The vertical dotted line indicates the point where either the sanction or no-sanction screen was revealed. The red and blue curves represent brain activities in the no-sanction condition and the sanction condition respectively[23-25, 32, 36].

These activation patterns are discussed further below. It is worthwhile to note here that our finding of Amygdala activation is consistent with recent evidence on its function. Although

Amygdala is typically associated with negative emotions and fear conditioning, emerging results suggest that Amygdala might be equally important to reward processing and goal directed behaviors. We elaborate this point below (see **Discussion**).

Neural activities predict trustee's repayment

We used standard parametric regression analysis to explore whether a trustee's neural activity at the revelation of the sanction screen might predict her subsequent backtransfer decision (which was made about 10 or 15 seconds later). Since the absolute backtransfer from a trustee does not inform a trustee's intention to cooperate, it is sensible to normalize the backtransfer by the maximum amount the trustee could have sent (the tripled investment amount). The backtransfer to tripled transfer amount ratio is a useful measure of a trustee's willingness to cooperate.

Our analysis revealed a brain area at the superior frontal gyrus (Fig. 4*A*, peak activity at MNI [24 52 20], $p < 0.005$, uncorrected). The activity of this area is negatively correlated with the backtransfer to investment amount ratio. Further ROI analysis shows that as this backtransfer ratio increases the BOLD signal at the DLPFC area decreases, and returns to the baseline level when the trustee fully cooperates (bottom panel, Fig. 4*A*, vertical bars indicate one s.e. of the mean). Positive parametric regression analysis identified several brain areas, including the medial frontal gyrus[39], the inferior frontal cortex, the middle temporal cortex, and the occipital cortex (Fig. 4*B*, SI Table 4**,** $p < 0.005$, uncorrected). Interestingly, one of those brain areas, the area in the ventromedial prefrontal cortex (Fig. 4*B*, peak activity at MNI [-4 56 -4], pink) significantly overlaps with the VMPFC region identified at the previous no-sanction – sanction contrast (Fig. 4*B*, yellow). The overlapping area is depicted in orange (Fig. 4*B*).

The ROI analysis (Fig. 4*B*, bottom panel) demonstrates this unique pattern of VMPFC activation. Although the VMPFC activity correlates with the repayment ratio in general, further separation of the VMPFC BOLD signal into sanction and no-sanction categories reveals a shift of the BOLD signal in both conditions (Fig. 4*B*, sanction in blue and no-sanction situation in red). Moreover, there is only weak evidence that the slope coefficients are different from one another (two-sample t-test, $p = 0.1$); the intercepts, however, are significantly different ($p < 0.01$, t-test). It is also interesting to note that, when the trustee plans to completely defect in the no-sanction situation, VMPFC activity remains at baseline. In contrast, when the trustee plans to defect under the sanction condition, VMPFC activity is well below baseline ($p < 0.05$, t-test). The fact

that brain activity at the VMFPC precedes the trustee's actual repayment choice by 10 to 15 seconds suggests that this brain area might be heavily involved in the trustee's final decision-making, and it might generate a BOLD signal predicting the trustee's repayment ratio. This signal is thus responsive in that it is susceptible to social cues (whether trustee is threatened by sanctions), as well as acting as predictive signal in that it parametrically modulates the trustee's final repayment.

Exploring the "once commodity, always commodity" hypothesis

Previous research suggests the "once commodity, always commodity" hypothesis that perception shifts can persist even when the source of the shift is removed[4]. We investigated, both behaviorally and at the neural level, whether being exposed to a sanction creates a perception shift that persists in future exchanges that do not include a sanction. To do this we focus on the 33 pairs whose investors chose both sanction and no-sanction at least once during the ten rounds. We categorized each round of each pair in one of three mutually-exclusive ways: 1) non-sanction trials before the investors imposed sanctions for the first time; 2) sanction trials; and 3) non-sanction trials experienced subsequent to sanction trials. In total, we obtained 88 observations on 20 unique subjects in group 1; 163 observations from 33 subjects in group 2; and 83 observations from 26 subjects in group 3. The "once commodity, always commodity" hypothesis predicts that Groups 2 and 3 should exhibit similar return behavior, and that Group 1 should return more on average than both the others.

We find trustees' returns (measured as percentage of tripled investment) are higher in Group 1 (49.2%) than in Group 2 (42.6%), and the decrease is (marginally) significant ($p = 0.10$, t-test, two-tailed). However, the back-transfer rate in Group 2 (42.6%) is less than found in Group 3 (51.6%), and the difference is statistically significant ($p = 0.03$, t-test, two-tailed). This is inconsistent with the hypothesis, and thus we do not find behavioral evidence supporting "once-commodity, always-commodity" in our environment.

To explore the hypothesis at the neural level, we conducted our imaging analysis using only the restricted sample of 33 subjects, and for only those observations that occurred in either Group 2 (sanction observations) or Group 3 (the no-sanction trials that occur subsequent to a sanction trial). The "once commodity, always commodity" hypothesis would predict neural activations consistent with "market" decision making in both groups. In fact, however, for the no-sanction - sanction contrast we found results again supporting social reward systems (VMPFC, PCC, OFC

and amygdala), but at a lower threshold (due to the substantially reduced sample size we used p = 0.01). SI Fig. 4 (Supplemental Information) reports the results of this analysis, and shows that the activations are closely related to those we discovered using the full sample. Similarly, we investigated the sanction - no-sanction contrast with the restricted sample. Again at a lower threshold (p = 0.01) we find activations in LIP that line up well with our original findings. It follows that neither our behavioral nor neural evidence supports the "once commodity, always commodity" hypothesis.

Discussion

Using an iterated version of the trust game with a sanction component we demonstrated an aversive effect of sanctions on human cooperation as measured by trustee's repayment in the investment game[5] (Fig. 2*B*). Recent theories that incorporate other regarding preferences, particularly inequality aversion or kindness, shed light on motives for trustees' decisions in standard trust games[6, 42-50]. These frameworks, however, cannot explain the detrimental effect of punishment on reciprocity. We hypothesized that this effect might owe to a "perception" shift, and supported this hypothesis using psychological theories of cognitive dissonance. Our data provide both behavioral as well as neural evidence supporting the perception shift hypothesis.

Differential brain activities in the No-Sanction – Sanction Contrast

Our perception shift hypothesis suggests that trustees not threatened with sanctions make their reciprocity decision within a social context and are directed by social norms. Indeed, when a trustee learns s/he has not been threatened with sanctions, we discovered activation of a neural network including the ventromedial prefrontal cortex (VMPFC), the right amygdala, the lateral orbitofrontal cortex (LOFC), and the posterior cingulated cortex (PCC). Activation of these reward-related pathways support our hypothesis for several reasons. One is that recent studies find elevated brain activity in the lateral OFC area when subjects choose to comply with social norms[51, 52], while the medial part of OFC (VMPFC) may be involved in preference generation and final decision-making[17, 31, 34, 53-55]. A second reason is that activation of this network might indicate the role of positive social emotions that can arise if a trustee interprets an investor's decision not to sanction as a benevolent gesture[5, 9, 51].

Regarding amygdala, although its activation in humans has been associated with negative emotions and fear conditioning, emerging evidence suggests amygdala might be equally important to reward processing[22, 53, 54, 56-60]. Also, reciprocal connections between the amygdala and the orbitofrontal cortex have been studied extensively, and the functional interaction between these two regions is thought to be essential in goal-directed behaviors[54,55,57-60]. Thus, increased amygdala activation in the absence of sanction threats, and in the context of the broader network of activations, supports the perception shift hypothesis in that it points to a neural network heavily involved in salient social signal processing as well as primary reward processing, evaluation, and final decision making.

Lastly, with respect to PCC, previous research has found that activity in this area might represent subjective preferences and engender impulsive behavior[61-63]. Greater activation of this brain region when a trustee receives a "no-sanction" signal might indicate the trustee is relatively more likely to send a significantly higher repayment (Fig. 2*B*) because their preferences shift in relation to cases where they receive sanction threats.

Differential brain activation in the Sanction – No-Sanction Contrast

The sanction – no-sanction contrast did not reveal any differential brain responses in the prefrontal cortex. Instead, we observed bi-lateral parietal cortex activation (SI Table 3). Parietal activity has been linked to the representation of expected utility in primate research and "rational" choices in both primates and humans[16, 63]. The fact that we do not observe differential activation of social or emotional systems under sanction threats seems to cast some doubt on the role of negative "intentions" in affecting behavior in this environment. Rather, this finding provides convergent support for the "cognitive shift" hypothesis that credible threats of sanctions generate a "cognitive shift" that diminishes social motivations and increases the likelihood of market-oriented earnings maximizing behavior[5-8].

Evidence of VMPFC as a Neural Integrator

The "perception shift" hypothesis requires the presence of a neural integrator to evaluate and compare inputs from various neural networks. Such an integrator would be expected to produce a signal that reliably predicts subjects' decisions. VMPFC is anatomically and functionally well suited to play this role, in that it projects to several brain areas that are heavily involved in reward valuation, preference generation and decision-making (e.g., striatum, amygdala, hippocampus and parietal cortex) and also is known to have intense local connections with

lateral orbitofrontal cortex. In investigating whether its activation predicts decisions, we indeed discovered that the VMPFC's activity is positively correlated with trustees' repayment ratio in both the sanction and no-sanction conditions. The specific brain area, revealed by linear regression analysis using trustees' repayment ratio as independent regressors, overlaps with the VMPFC area previously identified using the no-sanction – sanction contrast (Fig. 4*B*). Furthermore, we performed a region of interest (ROI) analysis of the overlapped region of VMPFC. A simple linear fit of VMPFC activation on repayment amount in both sanction and no-sanction conditions indicates no statistically significant difference in the estimated slope coefficients between conditions, yet a statistically significant difference in intercepts (Fig. 3, Fig. 4*B*).

Our findings regarding the VMPFC echo earlier results where experimenters, using a different paradigm, reported data suggesting that activations in a neural network including the VMFPC positively reinforce reciprocal altruism[42]. Our study, however, is unique in that we not only showed that VMPFC activity predicts trustee's reciprocal decisions, but also demonstrated that the same area's activity is susceptible to emotionally salient social cues (in particular, sanction or absence of sanction). Taken together, these results may indicate a common ground for the neural representation and interaction of monetary and social rewards[26, 39, 59, 60].

It is worth noting that we also found neural activity in the dorsal lateral prefrontal cortex (DLPFC) to be negatively correlated with trustee's repayment ratio. This correlation is revealed by a whole brain multi-linear regression analysis (Fig. 4*A*). While the DLPFC seems physiologically poorly positioned to play the role of integrator, the fact that the DLPFC response declines as the repayment ratio increases supports the idea that the DLPFC is important in cognitive control. It corroborates the view that the DLPFC represents goals and the means to achieve them in goal-directed behaviors[64-66]. For example, in an ultimatum game study using rTMS, DLPFC's activity seems to be crucial in implementing goal-directed behavior by overriding conflicting impulses[67]. Further, our findings that DLPFC response remains at baseline when a trustee cooperates perhaps suggests that conflict between altruistic and selfish responses require other brain networks to dynamically modulate the DLPFC's activity.

Conclusion

When not threatened with sanctions, trustees participating in a money exchange game displayed consistent activation patterns in a brain network previously linked to social reward

processing and decision-making[68, 69] These brain areas include LOFC, VMPFC, amygdala, and DLPFC. Brain activity in the same VMPFC area was found to correlate with the trustee's repayment ratio, which can be interpreted as a metric for cooperation. The presence of a sanction threat diminished activity in the social reward network but resulted in significantly increased activity in parietal cortex, an area implicated in rational cost-benefit analysis and decision making in humans. The sanction/no-sanction signal influences trustees' cooperation by modulating baseline activity of the VMPFC while leaving other parameters (such as correlation) unchanged.

Our particular focus was sanction effects: our task was not directly designed to probe the neural correlates of trust and reciprocity. Indeed, we did not find differential activations in some regions previously implicated in trust and reciprocity decisions[27, 28]. For example, caudate activity is known to be a reliable predictor of trustees' "intentions to trust"[27]. Evidence also suggests that putamen activity encodes efficiency, that the activation of insular can be mapped to inequity, and that caudate encodes a unified measure of efficiency and inequity. We find the activities of striatum, insula and anterior cingulated cortex not to vary across sanction-no sanction contrasts. Because one would perhaps expect differential activation of these areas if negative "intentions" were a driving force for sanctions' detrimental effects, the absence of differential activations seems to offer convergent evidence supporting the "perception shift" hypothesis.

An important issue left open by our study is to understand why investors choose to use ineffective sanctions. It seems unlikely that the reason is to enforce high backtransfer requests. In particular, across investment levels, request amounts do not differ between the sanction and no-sanction conditions (Fig. 2*A*). It is also unlikely that investors use sanctions to reduce potential inequality, since differences between players' payoffs are not statistically significantly different between the sanction and no-sanction conditions (two-sample t test, $p = 0.135$). Investors might use sanctions to encourage trustees not to defect. However, such a preemptive tactic by the investor is greeted with significantly less trustee cooperation (Fig. 2*B*). This finding is consistent with previous studies revealing non-monotonic effects of incentives on behavior[5-9]. We leave further investigation of investors' sanction decisions to future research.

In sum, our data are consistent with the view that detrimental effects of sanctions on human altruism can be explained by a "perception shift" that leads one to become more self-interested in market-environments that include prices (sanctions). While further work needs be done, this

result seems to cast some doubt on the role of negative "intentions" in reducing reciprocal tendencies in our specific investigation. More broadly, our findings advance understanding of the neural mechanisms underpinning human kindness and selfishness in social environments.

Methods

Subjects. Healthy subjects ages18-58 (N = 104, 61 females, age mean ± s.e 28.2±0.7) participated in the task. Half (52) of the subjects were randomly assigned as investors and the other half as trustees. 52 investors ages 20-58 (36 females, age mean ± s.e 31.1±1.2) and 52 trustees ages 18-35 (25 females, age mean ± s.e 25.4±0.4). Subjects were with normal or corrected vision and without any past or current neurological or psychiatric conditions, or structural brain abnormalities. All subjects were recruited through advertisements in local newspapers and internal school flyers. Informed consent was obtained using consent from approved by the Baylor College of Medicine Institutional Review Board.

Experiment. Subjects lay supine with their heads in the scanner bore and observed the rear-projected computer screen via a 45º mirror mounted above subjects' faces on the head coil. Subjects' choices were registered using two MRI-compatible button boxes.

Data Analysis.

Image acquisition and pre-processing:
High-resolution T1-weighted scans (1x1x1 mm) were acquired on Siemens 3T Allegra scanners using an MRPage sequence (Siemens). Functional images details: echo-planar imaging; repetition time (TR) = 2000 ms; echo time (TE) = 40 ms; flip angle = 90º; 64x64 matrix with 26 4mm thick axial slices, yielding functional 3.4x3.4x4 mm^3 voxels. To optimize functional sensitivity in the orbital frontal cortex (OFC), we acquired images using an oblique 30º to the AC-PC axis. All the imaging data was processed and analyzed using SPM2 (http://www.fil.ion.ucl.ac.uk/spm/software/spm2) and xjView (http://people.hnl.bcm.tmc.edu/cuixu/xjView). Functional images were realigned using a six-parameter rigid-body transformation. Each individual's structural T1 image was co-registered to the average of the motion-corrected images using 12-parameter affine transformation. Individual T1 structural images were segmented into grey matter, white matter, and csf before the

individual grey matter was nonlinearly warped into MNI grey matter template. Functional images were then slice-timing artifact corrected, normalized into MNI space by applying the transformation matrix adopted from previous T1 warping. Images were then smoothed with an 8 mm isotropic Gaussian kernel and high-pass filtered in the temporal domain (filter width of 128s).

Statistical Analysis:

General linear model (GLM) analysis: Functional images were divided into separate rounds (10 rounds) that included all images preceding each round by 20 seconds and following the end of each round by 8 seconds. Separate general linear models were specified and estimated for each round of the task for each subject. All visual cues and motor responses were constructed and estimated independently for each subject by convolving a delta function at the onset of those events with a canonical hemodynamic response function implemented within SPM2[70].

The random effect depicted in Figure 3 and Supplementary Table 2 and 3 were performed as the following: Fixed-effect analysis was performed for each round for each subject to estimate the brain activity of effects of interest. Beta images generated from above analysis were further separated into two uneven groups: 281 (no-sanction condition) and 238 (sanction condition) contrast images of a single between-group factor (sanction or no-sanction) and a two-sample t-test was performed. Supplementary Table 2 identified brain regions with significant greater activity ($T_{517} = 3.11$, $p < 0.001$, uncorrected) in response to “no-sanction” screen relative to “sanction” screen. Supplementary Table 3 identified brain regions with significant greater activity ($T_{517} = 3.11$, $p < 0.001$, uncorrected) in response to “sanction” screen versus “no-sanction” screen.

Region of interest (ROI) analysis: ROI analysis for four brain regions in Figure 3 (VMPFC, LOPFC, Amygdala and DLPFC) were performed on the 5 most significantly activated voxels from the t-test, which was depicted in supplementary Table 2. The spatially averaged signal was linearly detrended within each round and time-locked to the display of “sanction/no-sanction” information to the trustee’s brain. The correlation between brain activity of VMPFC and DLPFC and the normalized repayment ratio illustrated in Figure 4 are based on averages grouped by the level of normalized repayment ratio ((amount of repayment)/(3xinvestment), binned into 5 normalized repayment ratio levels: [0-0.2), [0.2-0.4), [0.4-0.6), [0.6-0.8), [0.8-1.0]. Trial events

numbers for the 5 repayment ratio levels are [23, 28, 75, 146, 9] for no sanction condition and [75, 18, 36, 84, 30] for sanction condition. Brain activities at VMPFC and DLPFC in Figure 4 are the averages of peak hemodynamic activities (at 4-6s after event onsets) and two data points surrounding the peak.

Figure Legends

Fig. 1. Experiment task. The task involves two subjects sequentially exchanging monetary units. Investors' choices are labeled in red and trustees' decisions in blue. (*A*) The investor makes three decisions sequentially: investment amount, backtransfer request and whether to threaten sanctions). Following this, the trustee makes the backtransfer decision. (*B*) Experiment timing. After each player makes her decision the results are displayed simultaneously to both subjects. A total of ten rounds are played and at the end of each round each player's earnings are revealed to both [see supporting information (SI) Fig. 1 & 2 for additional details].

Fig. 2. Summary of players' decisions when sanctions are threatened vs. not threatened (error bars are S.E.M.). (*A*) The investor's request as a function of the investment amount. The dotted line indicates a request of 2/3 of the tripled investment amount, which implies equal earnings for investor and trustee. The blue and red curves indicate investors' requests under the threat and no-threat of sanctions conditions respectively. (*B*) The trustee's repayment as a function of investor's investment. The dotted line indicates a backtransfer amount of 2/3 the tripled investment, which implies equal earnings for the investor and trustee. The blue and red curves indicate trustee's backtransfer under the threat and no-threat of sanctions conditions respectively (see also supporting information (SI) Fig. 3.).

Fig. 3. The trustee's brain regions showing greater activation in the "no-sanction" condition than the "sanction" condition ($p<0.001$, uncorrected, cluster size $k > 5$ voxels). (*A*) A random effects general linear model analysis revealed several brain regions significantly more activated by the revelation of "no-sanction". These regions include ventromedial prefrontal cortex (peak activation MNI coordinate [4 56 -4]), right Amygdala (peak activation MNI coordinate [24 0 -20]), right lateral orbitofrontal cortex (LOFC, peak activation MNI coordinate [32 52 -4]) and posterior cingulate cortex (PCC, peak activation MNI coordinate [4 -24 36]). (*B*) Mean event-related time courses of four brain regions (dashed line indicates the time onset; error bars are S.E.M.). Bold signal changes in the VMPFC, LOFC, Amygdala and PCC are all significantly greater when the trustee is in the "no-sanction" condition (red traces) than when she is in the "sanction" condition (blue traces).

Fig. 4. Trustees' brain regions whose activations are parametrically correlated with trustees' normalized backtransfer (defined as the ratio of the backtransfer and the tripled investment amount). (*A*) Brain activity at dorsal lateral prefrontal cortex (DLPFC, peak activation MNI coordinate [24 52 20]) is negatively correlated with trustees' normalized backtransfers ($p<0.001$, uncorrected, cluster size $k > 5$ voxels). (*B*) A general linear model ($p<0.005$, uncorrected, cluster size $k > 5$ voxels) reveales that a subset of voxels (peak activation MNI coordinate [-4 56 -4]; magenta) in the VMPFC area (yellow, the overlap was labeled in orange) which was previously identified in Figure 3A strongly and positively predicts trustees' normalized backtransfers. Further region of interest (ROI) analysis indicates that the VMPFC's activity is correlated with trustees' normalized backtransfers in both the sanction and no-sanction conditions. The slopes of the two curves (red & blue) are not significantly different from one another ($p = 0.1$, t-test) while the intercept of the "no-sanction" curve (red) is significantly greater than the intercept of the "sanction" curve (blue, $p <0.01$, t-test).

References

1. Fehr, E. , Gächter, S. (2002). Altruistic Punishment in Humans (97.4KB,), *Nature* 415, 10: 137-140

2. Bolton P, Dewatripont M (2005) in *Contract Theory* (Cambridge, MIT Press), pp 688.

3. Camerer CF (2003) in *Behavioral Game Theory: Experiments in Strategic Interaction:The Roundtable Series in Behavioral Economics* (Princeton, Princeton University Press), pp 544.

4. Andreoni J, Harbaugh W, Vesterlund L (2003) The carrot or the stick: Rewards, punishments, and cooperation. *Am Econ Rev* 93: 893-902.

5. Houser D, Xiao E, McCabe K, Smith V (2008) When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior* 62(2): 509-532.

6. Gneezy U, Rustichini A (2000) A fine is a price. *Journal of Legal Studies* 29: 1-17.

7. Deci EL, Koestner RM, Ryan RA (1999) Meta-analytic review of experiments examining the effect of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* 125: 627–668.

8. Lepper M, GreeneD (1978) The hidden cost of reward in *New Perspectives on the Psychology of Human Motivation.* (John Wiley Press).

9. Fehr E, Rockenbach B (2003) Detrimental effects of sanctions on human altruism. *Nature* 422:137-40.

10. Dickinson D, Villeval M (2004) Does monitoring decrease work effort? The complementarity between agency and crowding-out theories. *IZA Discussion Papers* 1222.

11. Frey B (1993) Does monitoring increase work effort? The rivalry between trust and loyalty. *Econ Inquiry*. 31: 663-670.

12. Bewley T (1999) Why wages don't fall during a recession (Cambridge, Harvard University Press).

13. Frey BS (1998) in *Not Just for the Money: An Economic Theory of Personal Motivation* (Beacon Press), pp 168.

14. Dorris MC, Glimcher PW (2004) Activity in posterior parietal cortex is correlated with the subjective desireability of an action. *Neuron* 44: 365-378.

15. Glimcher PW (2002) Decisions, Decisions, Decisions: Choosing a Neurobiological Theory of Choice. *Neuron* 36: 323-332.

16. Platt ML, Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400: 233-238.

17. Glimcher PW, Rustichini A (2004) Neuroeconomics: the concilience of brain and decision. *Science* 306: 447-452.

18. Kringelbach ML (2005) The human orbitofrontal cortex: linking reward to hedonic experience. *Nat Rev Neurosci* 6: 691-702.

19. Seguin JR (2004) Neurocognitive elements of antisocial behavior: Relevance of an orbitofrontal cortex account. *Brain Cogn.* 55: 185-197.

20. Camille N, *et al*. (2004) The involvement of the orbitofrontal cortex in the experience of regret. *Science* 304: 1167-1170.

21. Adolphs R (2001) The neurobiology of social cognition. *Curr Opin Neurobiol* 11: 231-9.

22. Veit R, *et al*. (2002) Brain circuits involved in emotional learning in antisocial behavior and social phobia in humans. *Neurosci Lett* 328: 233-236.

23. Adolphs R (2003) Cognitive neuroscience of human social behaviour. *Nat Rev Neurosci* 4: 165-78.

24. Moll J, Zahn R, de Oliveira-Souza R, Krueger F, Grafman J (2005) Opinion: the neural basis of human moral cognition. *Nat Rev Neurosci* 6: 799-809.

25. Pellis SM, *et al*. (2006) The effects of orbital frontal cortex damage on the modulation of defensive responses by rats in playful and nonplayful social contexts. *Behav. Neurosci* 120: 72-84.

26. Kringelbach ML (2005) The human orbitofrontal cortex: linking reward to hedonic experience. *Nat Rev Neurosci* 6: 691-702.

27. King-Casas B et al (2005) Getting to know you: Reputation and trust in a two-person economic exchange. *Science* 308:78-83.

28. King-Casas B et al (2008) The rupture and repair of cooperation in borderline personality disorder. *Science* 321:806-810.

29. Fellows LK, Farah MJ (2003) Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain* 126: 1830-1837.

30. Aron AR, Robbins TW, Poldrack RA (2004) Inhibition and the right inferior frontal cortex. *Trends Cogn Sci* 8: 170-177.

31. Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF (2005) Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310: 1680-1683.

32. Volz KG, Schubotz RI, von Cramon DY (2006) Decision-making and the frontal lobes. *Current opinion in neurology* 19: 401-406.

33. Coccaro EF, McCloskey MS, Fitzgerald DA, Phan KL (2007) Amygdala and Orbitofrontal Reactivity to Social Threat in Individuals with Impulsive Aggression. *Biol Psychiatry* 62(2): 168-178.

34. Schaefer M, Rotte M (2007) Thinking on luxury or pragmatic brand products: Brain responses to different categories of culturally based brands. *Brain Research* 1165: 98-104.

35. Bechara A, Damasio H, Tranel D, Damasio AR (1997) Deciding advantageously before knowing the advantageous strategy. *Science* 275: 1293-1295.

36. Bechara A, Damasio H, Damasio AR, Lee GP (1999) Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *J Neurosci* 19: 5473-5481.

37. Damasio A (2006) in *Descartes' Error* (VINTAGE RAND), pp 352.

38. Kringelbach ML, Rolls ET (2004) The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Prog Neurobiol* 72: 41-72.

39. McClure SM, *et al*. (2004) Neural correlates of behavioral preference for culturally familiar drinks. *Neuron* 44: 379-387.

40. Friedman, J. (1985) Cooperative equilibria in finite horizon noncooperative supergames. *Journal of Economic Theory.*35. 390-398.

41. Benoit, JP. Krishna, V. (1985) Finitely Repeated Games. *Econometrica*, 53(4). 905-922

42. Rilling J, *et al*. (2002) A neural basis for social cooperation. *Neuron* 35: 395-405.

43. Falk A, Fischbacher U (2006) A theory of reciprocity. *Game Econ Behav* 54: 293-315.

44. Falk A, Fehr E, Fischbacher U (2003) On the nature of fair behavior. *Economic Inquiry* 41: 20-26.

45. Cox J (2004) How to identify trust and reciprocity. *Game Econ Behav* 46: 260-281.

46. Engelmann D, Strobel M (2004) Inequity aversion, efficiency and maximin preference in simple distribution experiments. *Am Econ Rev* 94: 857-869.

47. Rabin M (1993) Incorporating fairness into game theory and economics. *Am Econ Rev* 83: 1281-1302.

48. Fehr E, Schmidt K (1999) A theory of fairness, competition and cooperation. *Q J Econ* 114: 817-868.

49. Dufwenberg M, Kirchsteiger G (2004) A theory of sequential reciprocity. *Game Econ Behav* 47: 268-298.

50. Ellingsen T, Johannesson M (2008) Pride and prejudice: The human side of incentive theory. *Am Econ Rev* 98(3): 990-1008.

51. Montague PR, Lohrenz T (2007) To detect and correct: norm violations and their enforcement. *Neuron* 56: 14-18.

52. Spitzer M, Fischbacher U, Herrnberger B, Gron G, Fehr E (2007) The neural signature of social norm compliance. *Neuron* 56: 185-196.

53. Gottfried JA, O'Doherty J, Dolan RJ (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301: 1104-1107.

54. Winstanley CA, Theobald DE, Cardinal RN, Robbins TW (2004) Contrasting roles of basolateral amygdala and orbitofrontal cortex in impulsive choice. *J Neurosci* 24: 4718-4722.

55. Arana FS, *et al*. (2003) Dissociable contributions of the human amygdala and orbitofrontal cortex to incentive motivation and goal selection. *J Neurosci* 23: 9632-9638.

56. Schultz W (2000) Multiple reward signals in the brain. *Nat Rev Neurosci* 1: 199-207.

57. Baxter MG, Parker A, Lindner CC, Izquierdo AD, Murray EA (2000) Control of response selection by reinforcer value requires interaction of amygdala and orbital prefrontal cortex. *J Neurosci* 20: 4311-4319.

58. Moll J, *et al*. (2002) The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *J Neurosci* 22: 2730-2736.

59. Holland PC, Gallagher M (2004) Amygdala-frontal interactions and reward expectancy. *Current opinion in neurobiology* 14: 148-155.

60. Schoenbaum G, Chiba AA, Gallagher M (1998) Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature neuroscience* 1: 155-159.

61. McCoy AN, Crowley JC, Haghighian G, Dean HL, Platt ML (2003) Saccade reward signals in posterior cingulate cortex. *Neuron* 40: 1031-1040.

62. McCoy AN, Platt ML (2005) Risk-sensitive neurons in macaque posterior cingulate cortex. *Nat Neurosci.* 8: 1220-1227.

63. McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. *Science* 306: 503-507.

64. Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24: 167-202.

65. Li J, McClure SM, King-Casas B, Montague PR (2006) Policy adjustment in a dynamic economic game. *PLoS ONE* 1: e103.

66. 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.

67. Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E (2006) Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 314: 829-832.

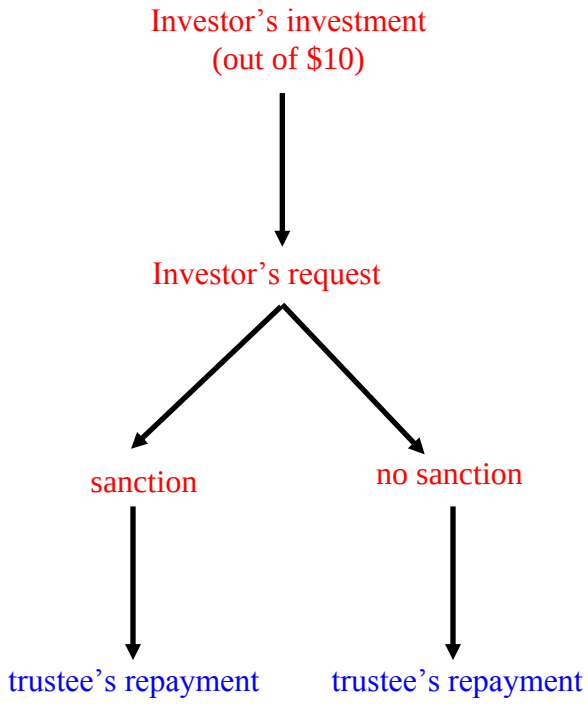
68. O'Doherty J, *et al*. (2003) Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia* 41: 147-155.

69. Rolls ET (2000) The orbitofrontal cortex and reward. *Cereb Cortex* 10 : 284-294.

70. Friston KJ, Frith CD, Frackowiak RS, Turner R (1995) Characterizing dynamic brain responses with fMRI: a multivariate approach. *NeuroImage* 2: 166-172.

Figure 1.

(A)


Investor's investment
(out of $10)
Investor's request
sanction
no sanction
trustee's repayment
trustee's repayment

(B)

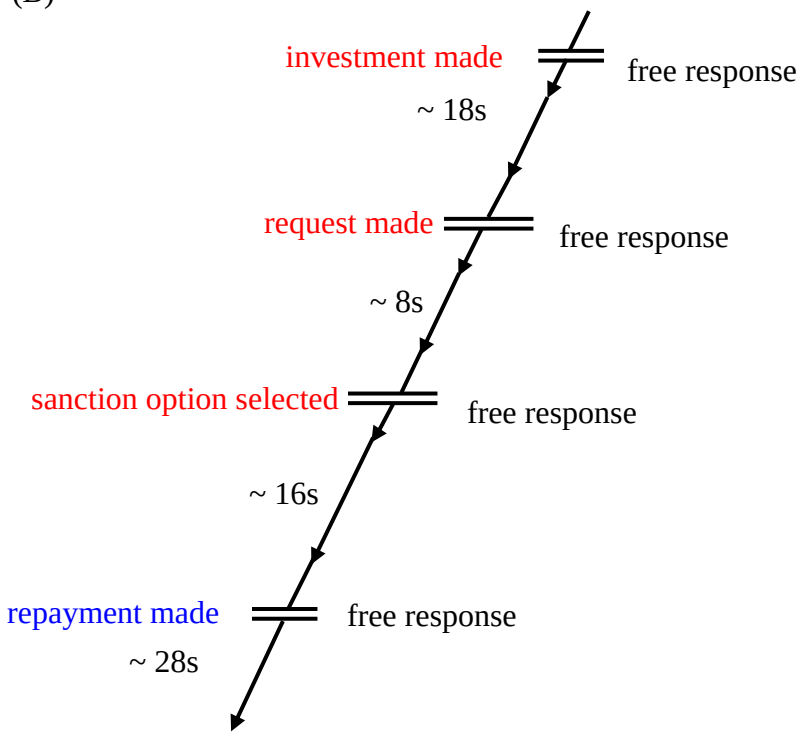
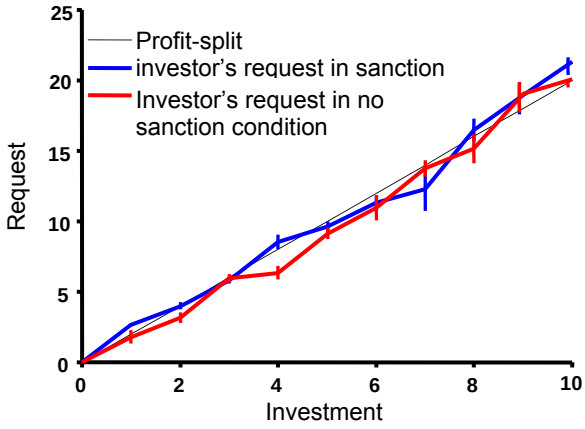

investment made
free response
~ 18s
request made
free response
~ 8s
sanction option selected
free response
~ 16s
repayment made
free response
~ 28s

Figure 2

(A)


Profit-split
investor's request in sanction
Investor's request in no sanction condition
Request
Investment
0
5
10
15
20
25
0
2
4
6
8
10

(B)

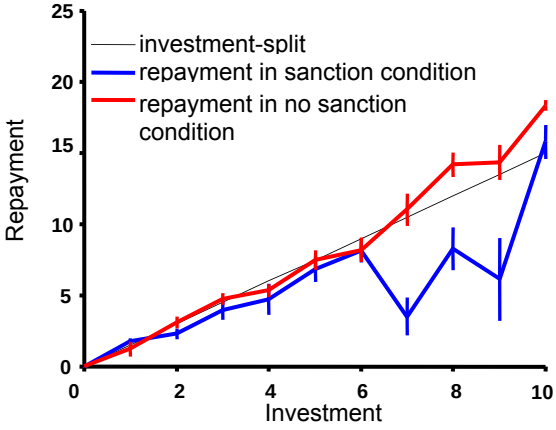
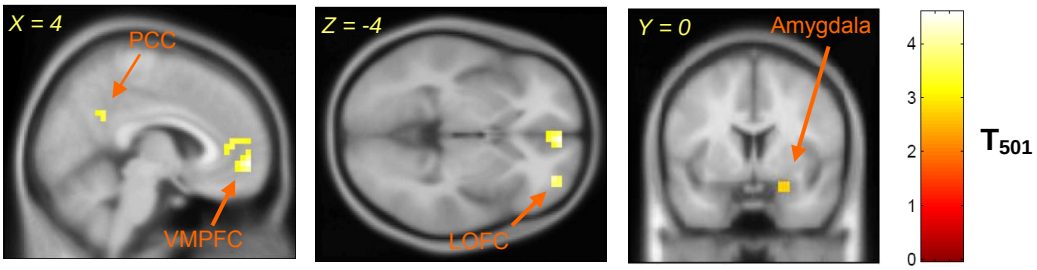

investment-split
repayment in sanction condition
repayment in no sanction condition
Repayment
Investment
0
5
10
15
20
25
0
2
4
6
8
10

Figure 3

(A)


X = 4
PCC
VMPFC
Z = -4
LOFC
Y = 0
Amygdala
4
3
2
1
0
T501

(B)

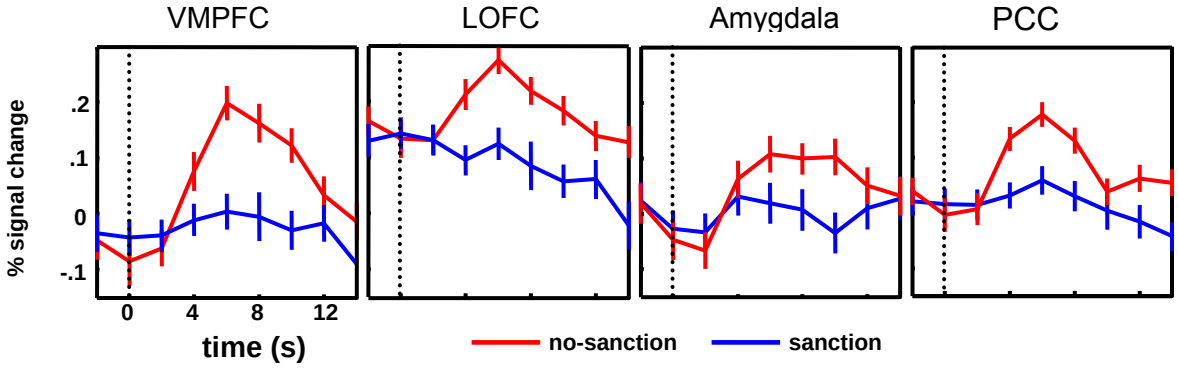
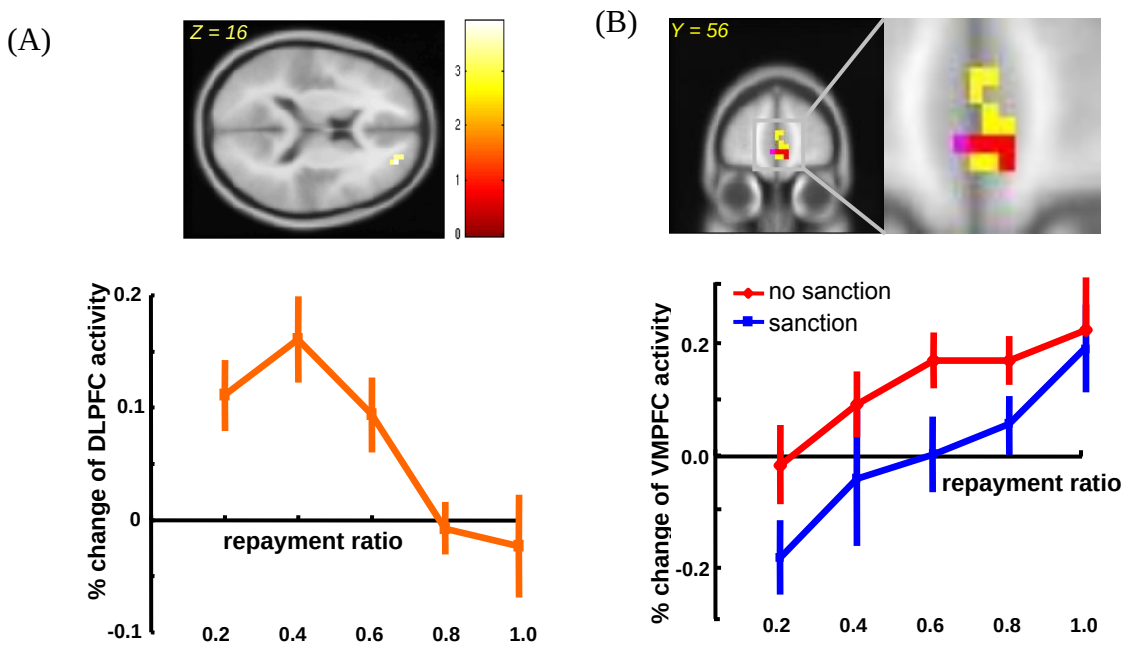

VMPFC
LOFC
Amygdala
PCC
% signal change
.2
.1
0
-.1
0
4
8
12
time (s)
no-sanction
sanction

Figure 4


(A)
Z = 16
3
2
1
0
% change of DLPFC activity
0.2
0.1
0
-0.1
repayment ratio
0.2
0.4
0.6
0.8
1.0
(B)
Y = 56
no sanction
sanction
% change of VMPFC activity
0.2
0.0
-0.2
repayment ratio
0.2
0.4
0.6
0.8
1.0

Supplementary Information for:

Neural Responses to Sanction Threats in Two-Party Economic Exchange

Jian Li[*†], Erte Xiao[‡], Daniel Houser[§] & P Read Montague[*¶]

March, 2009

[*]Department of Psychology, New York University, New York, NY 10003, USA. [‡]Department of Social and Decision Sciences, Carnegie Mellon University, USA. [§]Interdisciplinary Center for Economic Science (ICES), George Mason University, Fairfax, VA 22030, USA. [¶]Menninger Department of Psychiatry & Behavioral Sciences, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.

[†]Correspondence should be addressed to J.L. (lijian@nyu.edu). Department of Psychology, New York University, 6 Washington Place Room 873, New York, NY 10003

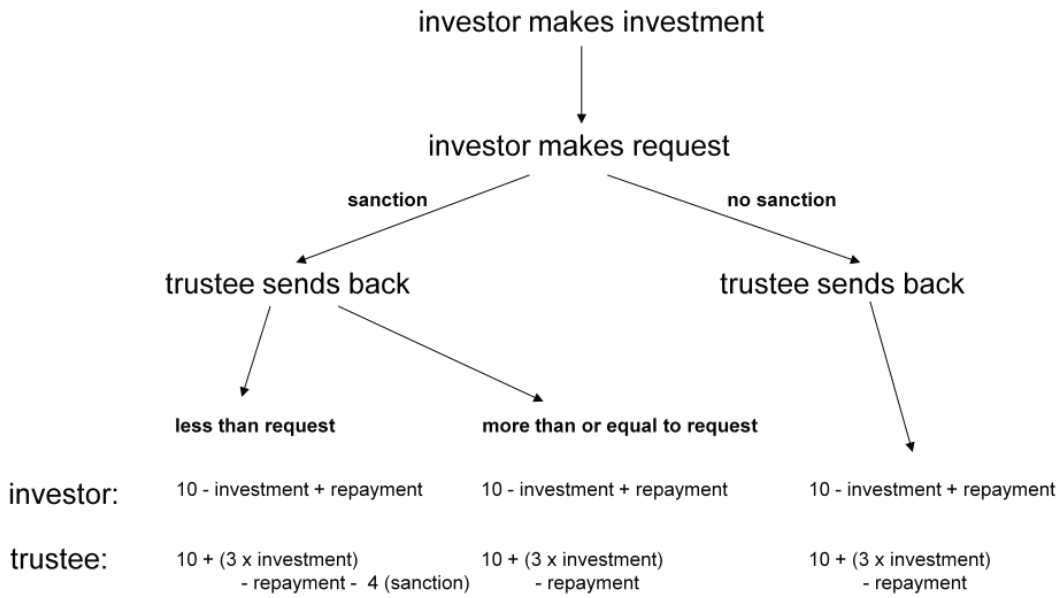
Author contributions: J.L., E.X., D.H., and P.R.M. designed research; J.L. performed research; J.L. and P.R.M. analyzed data; and J.L., E.X., D.H. and P.R.M wrote the paper.

Acknowledgements

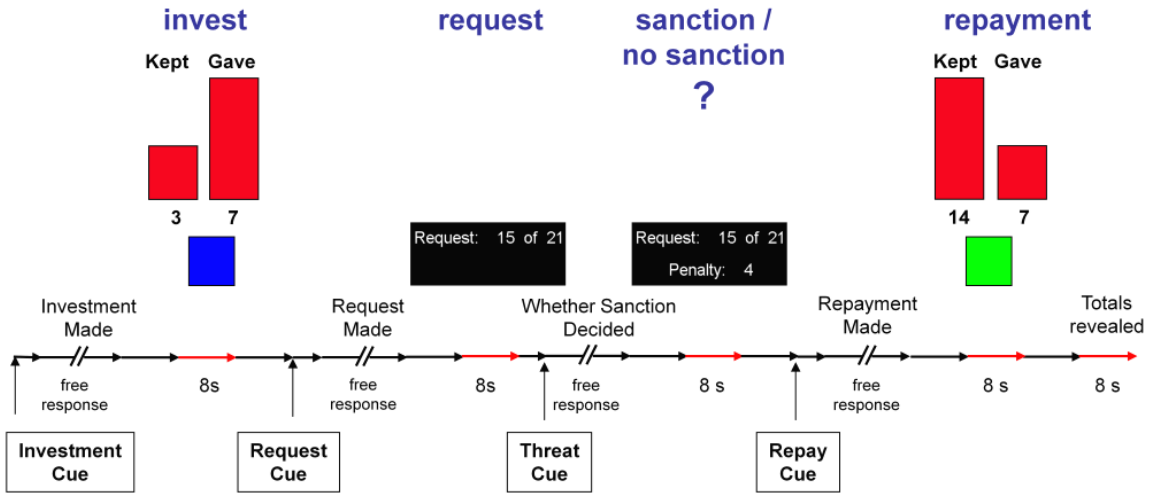
This research was supported by the Kane Family Foundation (P.R.M.), NINDS grant NS-045790 (P.R.M.) and NIDA grant DA-11723 (P.R.M.). We thank N. Apple for experimental design, S.M. McClure for helpful discussions, C. Bracero & J. McGee for fMRI images collection.

Competing interests statement

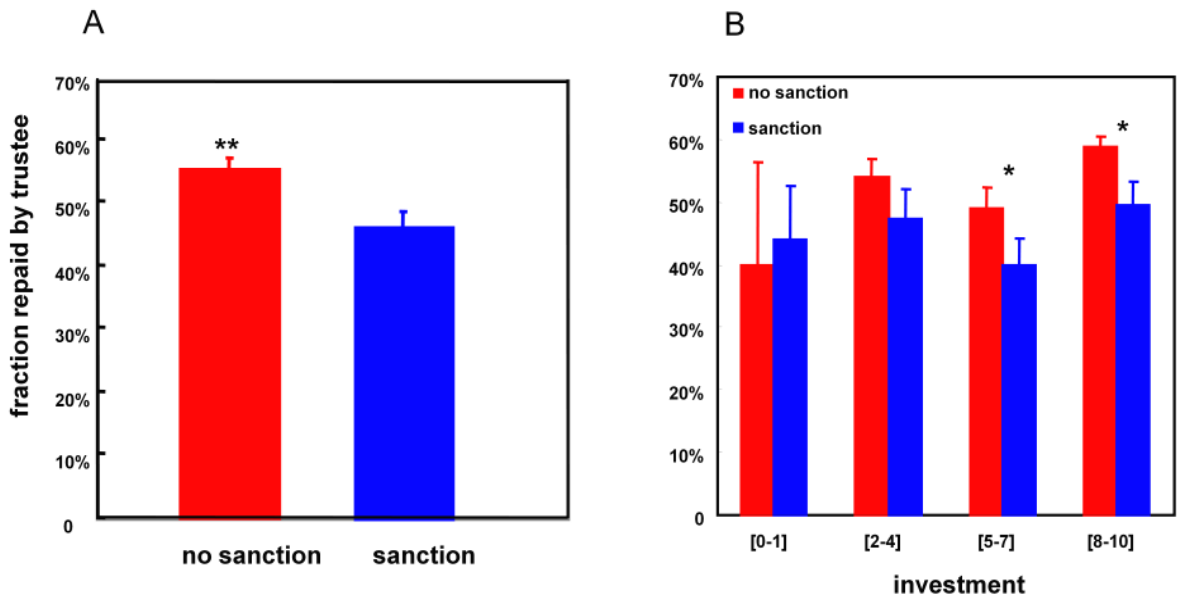
The authors declare no competing interests.



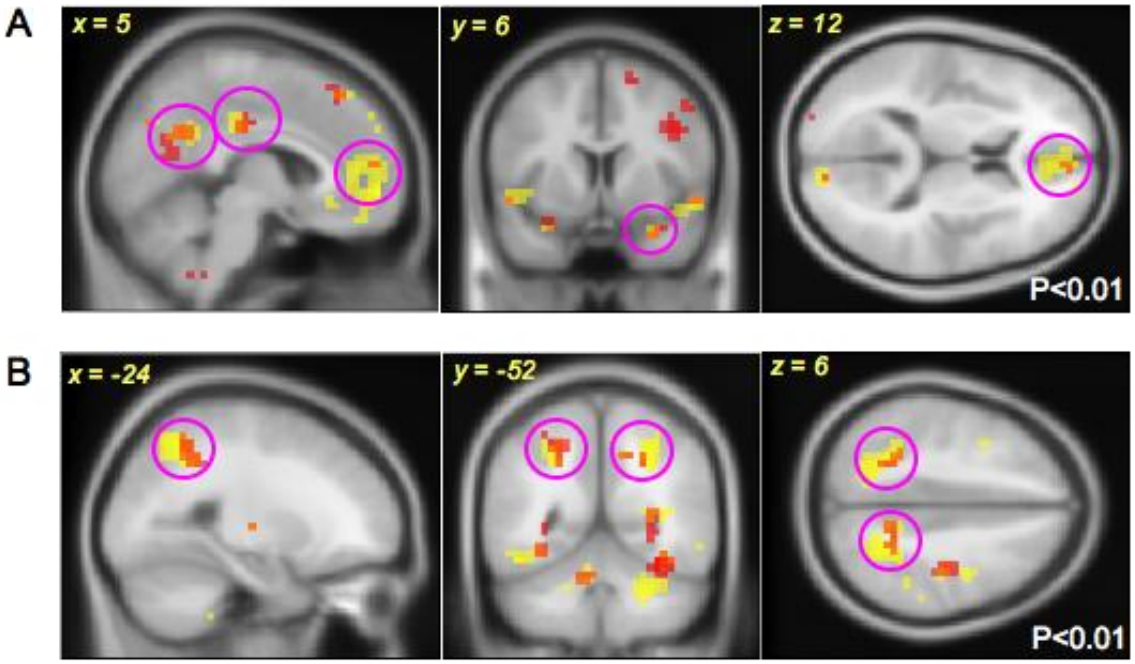
SI Figure 1. The two-player investment game. Two players are paired with each other anonymously. Both investors and trustees are endowed with 10 points at the beginning of each round of the experiment (10 rounds in total). The investor first decides how many points to invest, how many to request back and, whether to threaten punishment. The trustee observes these three pieces of information, and then decides how many points to send back to the investor. If the trustee returns less than the investor requested, and if the investor chose the threat option, then a penalty of four points is deducted from trustee's final earnings. If the threat was not chosen then trustees' and investors' earnings depend only on the amounts sent and returned, respectively, as described above.



SI Figure 2. Timeline for the two-player investment game. Each pair of subjects completed 10 consecutive exchanges. Each exchange began with a screen that indicated the beginning of the round, followed by a cue to invest. The investor then invested from 0 to 10 monetary units. After the investor's decision, the investment was displayed to both subjects for 8 seconds. The timing for the investor's next two decisions, the backtransfer request and whether to choose to threaten a sanction, proceeded in an identical manner. After the investor completed three decisions the trustee was prompted to return an amount (from 0 to the tripled investment amount) back to the investor. The trustee's decision is revealed to both subjects for 8 seconds followed by 8 seconds of a blank screen. That round's total earnings for both subjects is then displayed. Rounds were separated by a variable 12- to 42 second interval.



SI Figure 3. Behavioral summary of trustee's repayment under threat and no threat situations. (A) Under the threat condition, trustees repay significantly less (as fraction of available points) to the investor ($p<0.01$). (B) This conclusion holds when all the trials are divided according to investment level. The repayment difference between no threat and threat conditions is significant at higher investment levels.



SI Figure 4. Direct comparison between brain activation revealed by full dataset (52 subjects) and "Once a commodity, always a commodity (OCAC) hypothesis" (33 subjects). (A) Brain areas such as PCC, VMPFC, Amygdala (labeled in magenta circles), as revealed by no-sanction vs. sanction contrast, are represented in the overlapping pattern (orange) from full 52 subject dataset (yellow) and restricted 33 subject dataset (red). (B) Bilateral parietal cortex (labeled in magenta circles), as revealed by sanction vs. no-sanction contrast, are represented in the overlapping pattern (orange) from full (yellow) and restricted (red) datasets.

Average behavior and payoff of investors and trustees

	Sanction	No-sanction	Significance
Investment	4.89	7.09	*
Request	10.06	13.89	--
Request/(3*Investment)	0.72	0.64	*
Repayment	6.05	12.04	--
Repayment/(3*Investment)	**0.46**	**0.55**	*
Repayment/Request	0.67	0.89	*
Investor's Payoff	**11.58**	**14.95**	*
Trustee's Payoff	17.01	19.22	--

* Indicates statistically significant

SI Table 1

Brain responses differentially activated in no-sanction vs. sanction situations

	peak MNI coordinates				
Region of activation	***X***	***Y***	***Z***	***voxels***	***Z***
non-threat – threat					
Medial Frontal Gyrus (R)	**4**	**56**	**-4**	**83**	**4.45**
Superior Temporal Gyrus (R)	**48**	**16**	**-16**	**52**	**4.52**
Superior Temporal Gyrus (L)	**-48**	**16**	**-12**	**31**	**3.76**
Lateral Frontal Gyrus (R)	**32**	**52**	**-4**	**15**	**4.03**
Superior Frontal Gyrus (R)	**20**	**40**	**48**	**35**	**3.78**
Superior Frontal Gyrus (L)	**-28**	**40**	**36**	**24**	**3.26**
Occipital Lobe (R)	**12**	**-92**	**12**	**12**	**3.07**
Occipital Lobe (L)	**-16**	**-88**	**-16**	**19**	**3.58**
Precuneus (R)	**4**	**-52**	**32**	**12**	**3.49**
Posterior Cingulate Cortex	**4**	**-24**	**36**	**11**	**3.41**
Inferior Frontal Gyrus (R)	**52**	**24**	**4**	**5**	**2.78**
Amgdala (R)	**24**	**0**	**-20**	**7**	**2.70**

Regions with 5 or greater significant voxels were identified using T-test, $p<0.005$(uncorrected).

SI Table 2

Brain responses differentially activated in sanction vs. no-sanction situations

	peak MNI coordinates				
Region of activation	***X***	***Y***	***Z***	***voxels***	***Z***
threat – non-threat					
Parietal Lobe (L)	**-24**	**-60**	**52**	**72**	**3.99**
Parietal Lobe (R)	**28**	**-48**	**40**	**81**	**4.13**
Inferior Temporal Gyrus	**-44**	**-68**	**-4**	**67**	**4.10**
Temporal Lobe	**28**	**-68**	**20**	**27**	**3.29**
Precentral Gyrus (R)	**44**	**-4**	**36**	**68**	**3.97**
Precentral Gyrus (L)	**-44**	**-8**	**36**	**80**	**3.79**
Fusiform Gyrus (R)	**36**	**-48**	**-16**	**18**	**3.63**
Medial Frontal Gyrus	**-8**	**-24**	**68**	**17**	**3.30**
Midbrain	**4**	**-12**	**-12**	**59**	**4.17**
Cerebellum	**24**	**-48**	**-36**	**44**	**4.19**

Regions with 5 or greater significant voxels were identified using T-test, $p<0.005$(uncorrected).

SI Table 3

Brain responses positively related to repay ratio by trustees

	peak MNI coordinates				
Region of activation	***X***	***Y***	***Z***	***voxels***	***Z***
Medial Frontal Gyrus	**-4**	**56**	**-4**	**6**	**2.84**
Inferior Frontal Gyrus	**36**	**16**	**-20**	**18**	**3.89**
Middle Temporal Gyrus	**-60**	**-60**	**8**	**5**	**3.42**
Temporal Lobe	**-52**	**-8**	**-28**	**7**	**3.40**
Occipital Lobe	**-16**	**-96**	**-8**	**9**	**3.37**

Regions with 5 or greater significant voxels were identified using T-test, $p<0.005$(uncorrected).

SI Table 4