Reputation for reciprocity engages the brain reward center

K. Luan Phan[a,b,1,2], Chandra Sekhar Sripada[a,1], Mike Angstadt[a], and Kevin McCabe[c]

[a]Department of Psychiatry and [b]Neuroscience Program, University of Michigan, Ann Arbor, MI 48109; and [c]Center for the Study of Neuroeconomics, George Mason University, Fairfax, VA 22030

Communicated by Vernon Smith, Chapman University, Orange, CA, June 8, 2010 (received for review March 5, 2010)

Brain reward circuitry, including ventral striatum and orbitofrontal cortex, has been independently implicated in preferences for fair and cooperative outcomes as well as learning of reputations. Using functional MRI (fMRI) and a "trust game" task involving iterative exchanges with fictive partners who acquire different reputations for reciprocity, we measured brain responses in 36 healthy adults when positive actions (entrust investment to partners) yield positive returns (reciprocity) and how these brain responses are modulated by partner reputation for repayment. Here we show that positive reciprocity robustly engages the ventral striatum and orbitofrontal cortex. Moreover, this signal of reciprocity in the ventral striatum appears selectively in response to partners who have consistently returned the investment (e.g., a reputation for reciprocity) and is absent for partners who lack a reputation for reciprocity. These findings elucidate a fundamental brain mechanism, via reward-related neural substrates, by which human cooperative relationships are initiated and sustained.

functional MRI | neuroeconomics | ventral striatum | cooperation | trust

Social scientists have long understood the importance of trust in social relationships. Economists have demonstrated that people are better off if they can build social capital by finding trustworthy partners (1, 2), whereas psychologists have emphasized failure to build trusting relationships can compromise well-being and lead to despair (3, 4). Social neuroeconomics is increasingly uncovering the brain mechanisms that explain the motivation to build social capital and sustain cooperative social relationships (5, 6). Previous studies have shown that even after controlling for material gain, exchanges that are fair and equitable between individuals engender increased activation in reward related-regions, including the ventral striatum (vSTR) and orbitofrontal cortex (OFC), and enhanced personal happiness (7). This human preference for cooperative outcomes also extends to preferences for social partners (5). Humans are highly effective at decoding the reputations of others on the basis of prior actions (8) and using this reputational information to bias subsequent emotional reactions (9) as well as decisions to cooperate (8, 10, 11). Moreover, stable patterns of mutual cooperation (11, 12), as well as encountering those with a reputation for cooperation elicits reward-related activation of the vSTR and OFC (9).

Although it is well known that vSTR and OFC code rewards (13, 14) and respond to cooperative outcomes (15, 16), it is not currently known whether the brain's response to reciprocal outcomes is modulated by one's partner's reputation for cooperative behavior learned through one's own interactions with that partner. The trust game (17) and other game theoretic approaches have been used previously to examine how reward circuit activity changes with experience of positive/negative outcomes with (10) and without (12) prior knowledge of moral character/reputation. However, no study has examined brain responses in a trust game in which reputation of partners was experimentally manipulated and participants decoded this reputation for cooperation in real time during fMRI scanning.

Using the trust game modified into an iterative format for functional MRI (fMRI), we tested two predictions: (*i*) Given evidence that vSTR and OFC respond to rewarding social (7, 18) as well as nonsocial (13, 19) outcomes, we predicted that these regions will show enhanced activation to positive outcomes involving reciprocation of one's trust; and (*ii*) on the basis of findings that vSTR and OFC track signals of stable mutual cooperation (11, 12) and differentially respond to acquired reputations for cooperation (9), we predicted that the vSTR and OFC signal will be enhanced to partners who have formed a cooperative reputation (but not with less reputable partners). However, reward-related regions including vSTR and OFC, as well the ventral tegmental area (VTA) of the midbrain, have also previously been shown to encode a prediction error signal (20–22), with enhanced activation in these regions observed during outcomes that deviate from expectations of reward. The prediction error model makes an opposed prediction regarding how vSTR, OFC, and VTA will respond on the basis of partner reputation. This model predicts that activation in these regions will be enhanced when positive outcomes occur with partners that have *not* developed a reputation for fairness (i.e., unfair partners), because a positive outcome with partners known to be unfair should be most unexpected.

In this study, 36 healthy adults engaged in a multiround, repeated interaction trust game involving a unique reputational manipulation (Fig. 1). Participants played the role of an investor [decision maker 1 (DM1)] who must decide whether to invest 20 monetary units (MU) to a trustee partner [decision maker 2 (DM2)]. If the investor chooses to keep the money, the money is evenly split and each person receives 10 MU with certainty. But if the investor chooses to invest the money, the money is doubled (40 MU) and the trustee can then choose to either "reciprocate" by sending back half the money to the investor, or "defect" by retaining the entire amount thereby sending nothing back to the participant (0 MU). A key manipulation was that investors played in repeated interactions with three different fictive partner types, each associated with different tendencies for reciprocity, which were unknown to the participant at the start of the experiment. Unbeknownst to the participants, the frequency of reciprocity was actually fixed at the following frequencies: (*i*) FAIR partner = 75%; (*ii*) UNFAIR partner = 25%; and (*iii*) INDIFFERENT partner = 50%. Thus, this task manipulation forces participants to learn the tendencies of their partners to reciprocate (on the basis of prior interactions with that partner) to maximize personal gains. An additional COMPUTER partner was included, which did not require real-time learning of reputation (participants were told ahead of time this partner reciprocates 50% of the time), and served as a nonsocial control. In

Author contributions: K.L.P., M.A., and K.M. designed research; K.L.P. and M.A. performed research; K.L.P., C.S.S., and M.A. analyzed data; and K.L.P., C.S.S., and K.M. wrote the paper.

The authors declare no conflict of interest.

[1]K.L.P. and C.S.S. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: luan@umich.edu.

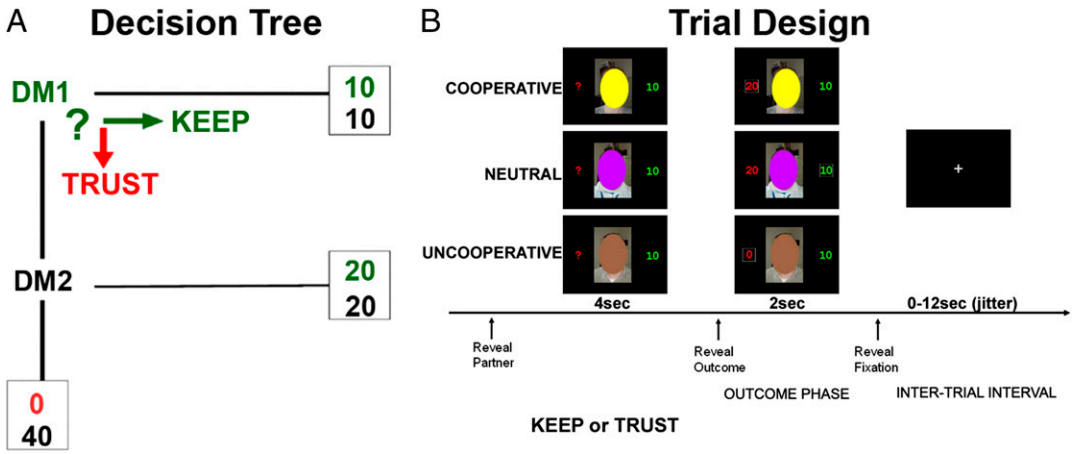

Fig. 1. (*A*) Decision tree. (*B*) Exemplar trial design showing three interactions with different partner types and their associated outcomes (e.g., yellow, COOPERATIVE partner and TRUST/Reciprocate outcome; purple, NEUTRAL partner and KEEP/{Defect} outcome; brown, UNCOOOPERATIVE partner and TRUST/Defect outcome).

sum, this study allowed us to measure both behavior and brain response to reciprocity (repayment following decisions to "trust") in relation to the reputation of human partners built over time.

Results

Investment Behavior in Relation to Partner Type. We expected that over time, participants would accurately associate a DM2 type with the corresponding likelihood for reciprocity and adjust their TRUST vs. KEEP choice accordingly. Behavioral results confirmed this prediction. A repeated-measures analysis of variance (ANOVA) using partner type (COOPERATIVE, UNCOOPERATIVE, NEUTRAL) and time (1–20) as within-subject factors showed a significant main effect of DM2 partner type on frequency (% of trials) of trust decisions [main effect of DM2 type: $F_{(2,70)} = 55.29$, $P < 0.001$]. Follow-up t tests revealed that participants correctly decoded each partner type's proclivity to reciprocate, as indicated by participants' differential investment behavior according to partner type (mean % invest was COOPERATIVE: 82.1% > NEUTRAL: 56.7% = COMPUTER: 53.9% > UNCOOPERATIVE: 29.2%) (all t tests: $P < 0.05$; Fig. 2*A*). Learning occurred rapidly (Fig. 2*B*), with differential investing based on partner type observed on average by the fifth trial, and stabilizing thereafter. The subjective rating of "trustworthiness" for each DM2 partner type was consistent with investment behavior, showing a significant main effect of DM2 partner type [$F_{(2,52)} = 30.12$, $P < 0.001$]; subsequent t tests revealed that subjects perceived COOPERATIVE (>NEUTRAL > UNCOOPERATIVE) partners as most trustworthy on the basis of subjective ratings collected after scanning (all t tests: $P < 0.05$; Fig. 2*C*).

Brain Response to Positive Feedback Following TRUST and KEEP Decisions. We were specifically interested in how the brain responds to outcomes after the participant has made the decision to TRUST and the decision to KEEP *separately*. In whole-brain neuroimaging analysis, we first examined trials in which participants trusted in their partner (TRUST trials) and contrasted instances in which their partner reciprocated against those in which their partner defected (TRUST/Reciprocate > TRUST/Defect), and as predicted, observed robust activations in bilateral vSTR [right: (20, 12, −10), $Z = 4.88$, $P < 0.05$ false discovery rate (FDR) corrected, 1,752 mm^3; left: (−26, 8, −8), $Z = 4.48$, $P < 0.05$ FDR corrected, 2,192 mm^3] (Fig. 3*A*), a region known to signal reward and pleasure (13, 19, 23). In addition to the vSTR, additional activations were observed in inferior occipital gyrus [(38, −90, −6), $Z = 5.14$, $P < 0.05$ FDR corrected, 12,896 mm^3], medial frontal gyrus/orbitofrontal cortex [(26, 36, 4), $Z = 4.87$, $P < 0.05$ FDR corrected, 8,056 mm^3], precentral gyrus [(32, −26, 58), $Z = 4.35$, $P < 0.05$ FDR corrected, 3,856 mm^3], and cerebellum [(42, −68, −50), $Z = 4.34$, $P < 0.05$ FDR corrected, 3,208 mm^3] (Fig. 3*A*). Visualization of the OFC cluster showed that it encompassed both gray and white matter, making functional interpretation problematic. Therefore, we parsed gray from white matter within this activated cluster using a canonical gray matter template (SPM5). From this procedure, we observed ≈34% of activated voxels falling within gray matter tissue with

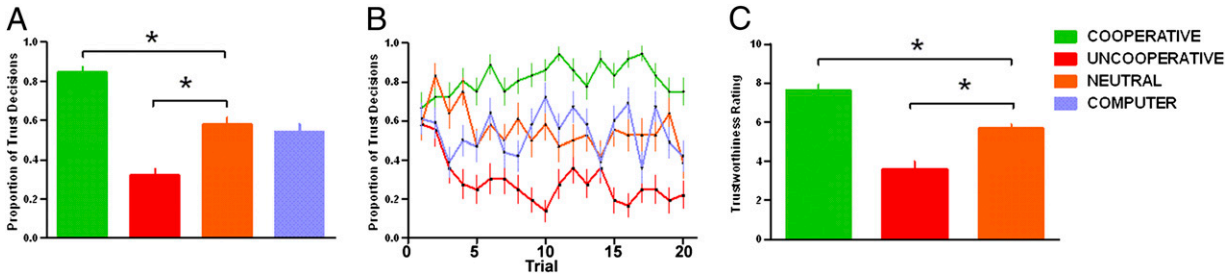

Fig. 2. Behavioral results. (*A*) Volunteers chose to trust COOPERATIVE more often than UNCOOPERATIVE, NEUTRAL, and COMPUTER partners (COOPERATIVE > NEUTRAL = COMPUTER > UNCOOPERATIVE; *$P < 0.05$). (*B*) Trial-to-trial trust behavior (proportion "trust" decisions collapsed across subjects) for each partner type over 20 trials during the fMRI experiment. (*C*) Volunteers perceived COOPERATIVE partners to be more "trustworthy" than UNCOOPERATIVE and NEUTRAL partners, based on subjective ratings collected after fMRI scan (COOPERATIVE > NEUTRAL > UNCOOPERATIVE; *$P < 0.05$).

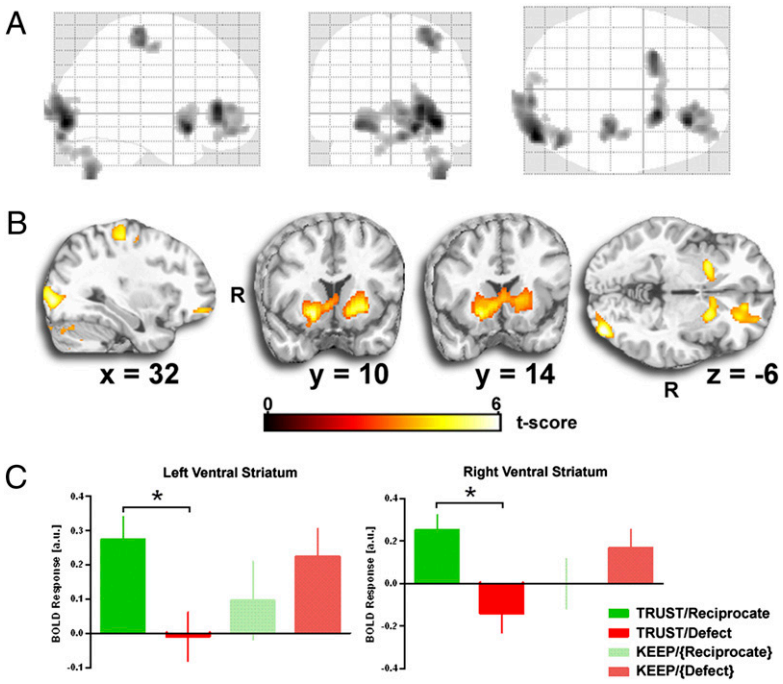

Fig. 3. Discrete and robust activation to positive reciprocity (TRUST/Reciprocate > TRUST/Defect contrast) of bilateral ventral striatum, right orbitofrontal cortex, precentral gyrus, inferior occipital gyrus, and cerebellum, displayed on a canonical glass brain (*A*) and canonical T1 brain template (*B*) (all activations are displayed at whole-brain voxelwise $P < 0.05$ FDR corrected). (*C*) Both left and right ventral striatum exhibits a positive response ("activation") to reciprocity following trust decisions (TRUST/Reciprocate) and a negative response ("deactivation") to defection following trust decisions (TRUST/Defect) (TRUST/Reciprocate > TRUST/Defect; *$P < 0.05$).

the location of the peak activation accordingly revised [(32, 52, −12), $Z = 3.95$, $P < 0.5$ FDR corrected, 2,720 mm^3]. In whole-brain search, we did not observe activation to positive feedback in VTA. Additionally, no significant activations to positive feedback were observed across the entire brain following KEEP decisions (e.g., Keep/{Reciprocate} > Keep/{Defect}).

We also analyzed parameter estimates of activation for each of the four different outcomes from functionally derived vSTR and OFC regions of interest (ROIs) to clarify the direction of activations; β weights were extracted from spherical ROIs [peak activations in the right vSTR (20, 12, −10), left vSTR (−26, 8, −8), and OFC (32, 52, −12)] for the TRUST/Reciprocate > TRUST/Defect contrast noted above in the whole-brain voxelwise search. Paired *t* tests of these parameter estimates revealed greater vSTR and OFC activation following TRUST trials in which their partner reciprocates compared with trials in which their partner defects [right vSTR: $t_{(35)} = 3.31$, $P = 0.002$; left vSTR: $t_{(35)} = 3.55$, $P = 0.001$; OFC: $t_{(35)} = 4.15$, $P < 0.001$] (Fig. 4*B*). As expected from the whole-brain analysis above, no difference was observed between KEEP/{Reciprocate} and KEEP/{Defect} trials [right vSTR: $t_{(35)} = 1.64$, $P = 0.110$; left vSTR: $t_{(35)} = 1.59$, $P = 0.122$; OFC: $t_{(35)} = -0.879$, $P = 0.385$] (Fig. 3*C*).

In sum, these results show robust vSTR and OFC responses to positive feedback (Reciprocates > Defect) following TRUST decisions (Fig. 3*B*) and that this activation is relatively specific to the TRUST/Reciprocate condition (Fig. 3*C*), suggesting that it represents a signal for actual reciprocity.

Brain Responses to Reciprocity by Partner Type. Given the pattern of bilateral vSTR as well as OFC activation to reciprocity noted above, we next examined how these responses are modulated by reputations for cooperation associated with different DM2 partner types; specifically, we extracted parameter estimates of activation for each of the four DM2 partner types from the spherical ROIs in vSTR and OFC described above centered at peak activations from the TRUST/Reciprocate > TRUST/Defect contrast in whole-brain voxelwise search. Results differed by region. In vSTR, a one-way ANOVA showed a significant main effect of partner type [right: $F_{(2,46)} = 4.87$, $P = 0.012$; left: $F_{(2,46)} = 6.89$,

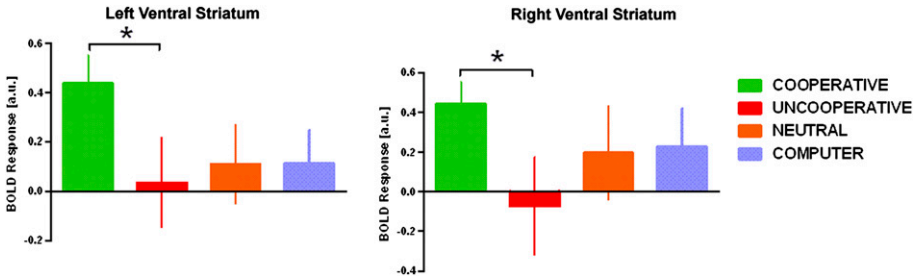

Fig. 4. Activation in both left and right ventral striatum in response to partner reciprocation compared with partner defection following TRUST decisions is selective for COOPERATIVE partners (COOPERATIVE > UNCOOPERATIVE; *$P < 0.05$).

$P = 0.002$]; follow-up t tests were conducted to compare activation between partner types (i.e., do the bars in Fig. 4 differ from each other?) and revealed greater vSTR activation to COOPERATIVE than to UNCOOPERATIVE partners bilaterally [right: $t_{(35)} = 2.41$, $P = 0.021$; left: $t_{(35)} = 2.63$, $P = 0.031$] (Fig. 4). Additionally, there was a trend toward greater activation in left vSTR to COOPERATIVE partners vs. NEUTRAL [$t_{(35)} = 1.68$, $P = 0.101$] and COMPUTER partners [$t_{(35)} = 1.89$, $P = 0.067$]. Paired sample t tests were performed to assess whether vSTR activation significantly differed between TRUST/Reciprocate vs. TRUST/Defect trials for each partner type (do the bars differ from zero?). Results showed that the difference was significant for COOPERATIVE partners *only* [right: $t_{(35)} = 3.94$, $P < 0.001$; left: $t_{(35)} = 3.84$, $P = 0.001$]. For the other partner types, vSTR responses to reciprocity did not significantly differ from responses to defection [right: UNCOOPERATIVE $t_{(35)} = -0.298$, $P = 0.768$; NEUTRAL $t_{(35)} = 0.840$, $P = 0.407$; COMPUTER $t_{(35)} = 1.19$, $P = 0.244$; left: UNCOOPERATIVE $t_{(35)} = 0.194$, $P = 0.847$; NEUTRAL $t_{(35)} = 0.691$, $P = 0.494$; COMPUTER $t_{(35)} = 0.842$, $P = 0.406$]. Moreover, follow-up one sample t tests revealed that enhanced activation in vSTR to COOPERATIVE partners in the TRUST/Reciprocate > TRUST/Defect contrast was driven by increased activation to reciprocation as opposed to decreased activation to defection, as ventral striatum robustly activated compared with fixation in TRUST/Reciprocate trials [right: $t_{(35)} = 3.346$, $P = 0.002$; left: $t_{(35)} = 4.042$, $P < 0.001$], but did not significantly differ from fixation in TRUST/Defect trials [right: $t_{(35)} = -1.284$, $P = 0.208$; left: $t_{(35)} = -0.294$, $P = 0.771$]. In OFC, a one-way ANOVA failed to show a significant main effect of partner type [$F_{(2,46)} = 1.25$, $P = 0.297$]. We also looked for activations outside of our a priori vSTR and OFC regions using an exploratory voxelwise whole-brain analysis for the TRUST/Reciprocate > TRUST/Defect contrast. A one-way ANOVA revealed no regions across entire brain that exhibited a significant main effect of partner type.

Discussion

We used event-related fMRI to investigate the neural correlates of reciprocity during iterative economic exchanges with fictive partners who develop different reputations for repaying (or not) the investment entrusted to them. The current work confirms previous results that reciprocity, as reflected in a positive return (a gain of 20 MUs) from investment, robustly engages reward-related regions including ventral striatum and orbitofrontal cortex. The key unique finding of this study is that the vSTR, but not the OFC, response is selective for partners who have consistently returned the investment (e.g., a "cooperative" reputation), and is absent for partners who lack a reputation for cooperation. Although Delgado and colleagues had previously shown that partner moral reputation, learned through written descriptions before fMRI scanning, modulates vSTR responding for neutral partners in the context of the trust game (10), the current study uniquely shows enhanced vSTR responding to cooperative partners whose reputation is ascertained through one's own real-time interactions during fMRI scanning (5).

We found enhanced activation in vSTR and OFC during trials in which participants trust their partners and their partner reciprocates compared with trials in which their partner defects. However, our design does not allow us to conclusively determine which among several possibilities explains this vSTR and OFC activation. In addition to the notion that vSTR and OFC activation represents a "reward" signal from social reciprocity, other interpretations are possible. Because the actual monetary payoff for TRUST/Reciprocate is the largest gain among all outcomes (gain of 20 MUs) and TRUST/Defect is the largest loss among all outcomes (loss of 10 MUs), vSTR and OFC activation could reflect the magnitude of this monetary difference. Also, it is interesting to note that vSTR and OFC also activate to the KEEP/{Defect} outcome whereas it shows no activation to KEEP/{Reciprocate} outcome (Fig. 3*B*). Both outcomes represent an actual payoff of 10 MUs; however, only the KEEP/{Defect} outcome signifies that a potential loss of 10 MUs was averted. Thus, it is also possible that vSTR and OFC activation represents a signal that the participant made a "good" decision; in other words, TRUST/Reciprocate and KEEP/{Defect} represent the two conditions in which the chosen action yielded a better outcome compared with the alternative action, and this might account for the fact that the vSTR activation was higher in these two cases.

Prior work has consistently found vSTR activity is reliably linked to the receipt of rewards, both concrete and abstract (13, 18, 19, 23, 24). For example, vSTR activity is demonstrated for receipt of primary rewards (e.g., squirts of juice) (13), monetary rewards (14), as well as social rewards (18). Moreover, activation in vSTR is closely tied to the subjective experience of positive emotions and pleasure (14). Our key unique finding is that in the context of social interactions, vSTR activity is significantly modulated by the reputation for cooperation of one's partner. In particular, in trials in which participants trust their partner, vSTR significantly responds to partner reciprocation compared with partner defection only for social partners that have a prior reputation for cooperative behavior. For partners that lack a cooperative reputation, vSTR responses to outcomes involving partner reciprocity do not significantly differ from its responses to partner defection. These results suggest that the value of social capital derived from interacting with trustworthy partners is "built into" the vSTR reward signal at a very basic level. In contrast to vSTR, another reward-related region, OFC, activated to reciprocation of trust (>defection) but did not respond selectively to reciprocation from cooperative partners. Rather this region activated to reciprocation from all partner types regardless of positive or negative reputation. This finding may be related to results from other studies that suggest that in contrast to vSTR, which robustly and primarily responds to positive stimuli (13, 14, 18) [note: others have posited that more dorsal regions of stratium respond to negative stimuli (25, 26)], OFC more reliably responds to reward-relevant stimuli of both positive and negative valence (27, 28).

Prediction-error models provide an alternative framework to explain activation in reward-related regions in the context of repeated social interactions involving monetary payoffs (11, 16, 29). According to these models, activation in regions such as vSTR, OFC, and VTA is enhanced by delivery of an unexpected reward and attenuated by omission of an expected reward (13, 20, 22). However, our observation in the current study of enhanced vSTR activity to COOPERATIVE compared with UNCOOPERATIVE and NEUTRAL partners cannot be fully explained by prediction-error effects. Our behavioral results of differential investment choices show that people were much more likely to invest in the COOPERATIVE partner than the other partner types, strongly suggesting they expected reciprocation from this partner type. Thus the outcome in which the COOPERATIVE partner reciprocates is likely to be more expected (rather than more novel or unexpected) than for UNCOOPERATIVE or INDIFFERENT partners. For this reason, enhanced vSTR activity to the COOPERATIVE partner cannot be explained by the unexpected nature of the reward, but rather appears to be driven by that partner's reputation for cooperation. However, consistent with findings from these prior studies (13, 20, 22), we did observe a small (although statistically nonsignificant) "deactivation" of vSTR response following the TRUST/Defect condition during which the participant expected reciprocity but encountered defection (thereby an omission of expected reward); the small number of trials within and across subjects in which these outcomes occurred to the COOPERATIVE partner precludes us from examining the effect of partner type on the extent of this vSTR deactivation. It is

noteworthy that another study of the trust game that used a reputational manipulation also failed to find effects consistent with the prediction-error model (10). This has led to the suggestion that the presence of reputational information may blunt or supersede prediction-error processing (16), a hypothesis that warrants further direct study.

It is worth discussing our findings in relation to those observed by Delgado and colleagues who had previously modified the trust game by informing participants (DM1s) about the "moral" character (e.g., good, neutral, bad) of fictitious partners via biographical sketches and examined how such reputations influenced trust behavior and brain response (10). Similar to our findings, the authors showed that the vSTR activates more to positive (gains), than to negative (losses), feedback. However, in contrast, reward-related responses were greatest for neutral partners and no differentiation (between gains and losses) was observed in vSTR reactivity to morally good partners (compared with neutral or bad partners). Delgado and colleagues argued that neutral partners elicit greater vSTR response because of enhanced need for reward-based learning, compared with partners with more certain reputations for either good or bad behavior. It is noteworthy that in the Delgado et al. study, participants learned about their fictive partners' reputation via explicit verbal descriptions, whereas in our study this socially relevant information was acquired over time via experience during the interactive trust game. Because learning from instruction and learning from experience may each engage distinct neural substrates (9), this methodological difference may help explain why Delgado and colleagues did not find enhanced activation in vSTR for good/fair partners. Interestingly, in the Delgado et al. study, despite explicit evidence that all three partners were reciprocating similarly, participants discounted the feedback information and continued to differentially invest in partners with a morally good reputation. In contrast, participants in this study utilized feedback information to guide their investment decisions.

Our findings indicate that studies of decision making and valuation in economic exchange paradigms should account for the role of reputation in modifying behavior and brain response. The brain's reward center reliably responds to food, money, and social rewards. But here we show that the brain's reward center selectively responds to monetary rewards received from partners with a reputation for cooperative play, but fails to respond to identical monetary rewards from partners who lack a reputation for cooperation. Economists are increasingly recognizing the importance of the formation of social capital for the success of individuals and societies. Our results show that ventral striatal reward signals robustly and selectively encode the value of gains realized from trustworthy partners, thus providing unique insights into the underlying brain mechanisms by which human social capital is produced and sustained.

Materials and Methods

Subjects. Thirty-six subjects (22 females; mean age and SD 30.03 ± 8.64 y) participated in the study. All were right-handed and healthy, without a history of psychiatric, neurologic, or major medical problems, and free of psychoactive medications at the time of the study. None of the subjects tested positive on a urine toxicology screen or alcohol breathalyzer on the day of scanning. All participants gave written informed consent for this study, as approved by the local institutional review board.

Trust Game Task. The fMRI task involved an event-related design (Fig. 1). Participants played the role of investor in a multiround ("repeated interaction") trust game against three anonymous partners, whose reputations for reciprocity had to be learned through interactions (20 trials with each partner) during the course of the game; in other words, to maximize monetary payoffs, the participants had to correctly link each partner with that partner's likelihood for reciprocity (reputation). Participants were instructed that they are assigned to be "decision maker 1" (DM1) in a "decision-making" game. They were also told that they would be playing with other, anonymous people who had previously participated in the same game as "decision maker 2" (DM2) and whose responses were previously recorded and now serve as DM2 "reactions" to their (DM1's) decisions. To make the scenario more credible, DM2s were represented by different face photographs but the "face" was obscured by a colored oval so as to convey a sense of anonymity. Any confounds of facial features, interpersonal attraction, personal identity, and emotional expression were reduced by this manipulation.

Subjects were further instructed to imagine they were playing the game with DM2 partners in real time. Participants were told that they could "win" as much as $20.00 based on cumulative outcomes of the experiment. As DM1, the participant was informed of the task in the following way (Fig. 1*A*):

(*i*) For each trial, you are given 20 monetary units (to be converted into actual money after the end of the experiment).
(*ii*) For each trial, you must make a decision to keep, and thus, equally divide the 20 units between yourself and your partner (KEEP) or to invest the 20 units (TRUST).
(*iii*) If you choose to KEEP then the actual outcome of this trial is complete (e.g., you will receive 10 units, and your partner will receive 10 units).
(*iv*) If you choose to TRUST, then the amount doubles to 40 monetary units, and the actual outcome of the trial is to be decided by DM2, who can choose to reciprocate by splitting the money equally with you (e.g., you will receive 20 units, and DM2 will receive 20 units) or defect by keeping the entire amount to himself/herself (e.g., you will receive 0 units, and DM2 will receive 40 units).

Participants were informed that they would play with three "types" of DM2 players classified on the basis of their previously recorded actions as (*i*) type 1: reciprocates > 50% of the time; (*ii*) type 2: reciprocates < 50% of the time; and (*iii*) type 3: reciprocates about 50% of the time. In addition, they were also told that they were playing with a computer (represented by an image of a desktop computer) that "reciprocates 50% of the time." Unbeknownst to the participants, the frequency to reciprocate an investment was actually fixed at the following frequencies: (*i*) type 1 = 75%; (*ii*) type 2 = 25%; and (*iii*) type 3 = 50%. These DM2 partner types are referred to here as COOPERATIVE, UNCOOPERATIVE, and NEUTRAL, respectively. Thus, the task manipulation of having participants repeatedly play three fictive partners forces participants to accurately ascertain the tendencies of their partners to reciprocate to maximize personal gains.

At the start of each trial (Fig. 1*B*), participants viewed one of three different obscured face photographs representing a DM2 type or they viewed an image of a computer. The color of the oval designated the type of DM2 and was counterbalanced across subjects, and participants were not aware of the mapping between color of oval and type of DM2 at the start of the experiment. The DM2/computer image appeared for 4 s during which the participants were instructed to make their choice (KEEP or TRUST) by button press. In real time and on the basis of the subject's own decision/choice, feedback was provided immediately in the form of a DM2/computer image reappearing for 2 s along with information about the participant's choice, as well as the DM2's actual (in instances of DM1 TRUST) or hypothetical (in instances of DM1 KEEP) response. This information was represented to the side of the DM2 image as the amount of money sent back to DM1 (either 0 or 20 monetary units, designating DM2's defect or reciprocate decision,respectively).

Each trial was separated by an intertrial interval (blank gray-scale screen with fixation crosshair), jittered from 0 to 12 s. There were a total of 80 trials equally representing the three types of DM2s and the computer (i.e., 20 trials of each), which were pseudorandomly presented and distributed evenly across four fMRI runs. After the experimental session was complete, participants were paid according to the actual outcomes accumulated over 80 trials of the task. In addition, subjects were debriefed after they completed a single postscan subjective rating questionnaire of "trustworthiness," one rating for each type of DM2 ("How much do you trust this person?") on a Likert scale of 1–10, anchored by the following descriptors (1, not at all trustworthy; 10, extremely trustworthy). Coupled with participants' investment behavior, this questionnaire allowed us to corroborate whether our reciprocity manipulation was successful in influencing participants' investment decisions *and* their perception of the different partner types.

Image Acquisition and Processing. Scanning was performed with BOLD (blood oxygenation-level dependent)-sensitive whole-brain fMRI on a 3.0 Tesla GE Signa System (General Electric) using a standard radiofrequency coil and associated software (LX 8.3, neuro-optimized gradients). Whole brain functional scans were acquired using a T2*-weighted reverse spiral sequence (echo time

= 25 ms, repetition time = 2,000 ms, 64 × 64 matrix, flip angle = 77°, field of view = 24 cm, 30 contiguous 5-mm axial slices per volume, aligned with the anterior commissure-posterior commissure line). A high-resolution T1 scan (3D-MPRAGE; repetition time = 25 ms; min echo time; field of view = 24 cm; slice thickness = 1.5 mm) was also acquired.

Data from all 36 participants met criteria for high quality and scan stability with minimum motion correction (<2-mm displacement) and were subsequently included in the data processing. Preprocessing steps were implemented using Statistical Parametric Mapping 5 software (SPM5; Wellcome Department of Cognitive Neurology, London, United Kingdom; www.fil.ion.ucl.ac.uk/spm). The first four volumes from each run were discarded to allow for T1 equilibration effects. Preprocessing followed conventional procedures: (*i*) slice time correction; (*ii*) spatial realignment; (*iii*) normalization to the Montreal Neurologic Institute (MNI) template through the use of nonlinear warping algorithm; (*iv*) spatial smoothing through the use of a Gaussian 8-mm full-width-half-maximum kernel; and (*v*) high-pass temporal filtering with a cutoff of 128 s to remove low-frequency drifts in signal. After preprocessing, statistical analyses were performed at the individual and group level using the general linear model (GLM) and Gaussian random field theory as implemented in SPM5 with regressors representing each partner type (COOPERATIVE, UNCOOPERATIVE, NEUTRAL, and COMPUTER) and the four types of outcomes: TRUST decisions in which one's partner *actually* reciprocates (TRUST/Reciprocate) or defects (TRUST/Defect) and KEEP decisions in which *hypothetically* one's partner would have reciprocated (KEEP/{Reciprocate}) or would have defected (KEEP/{Defect}); here, braces represent the DM2's hypothetical choice had the participant chosen to trust. Regressors of interest (condition effects) were generated using a canonical hemodynamic response function (HRF) corresponding to the onset of the outcome being revealed. In the first-level analysis, these regressors were convolved with the canonical HRF, using the temporal derivative to account for intersubject variability in BOLD signal time to peak. In the second-level analysis, subjects were treated as a random effect and images were thresholded using a voxelwise threshold of $P < 0.05$, FDR corrected for multiple comparisons across the entire brain (30).

Prior evidence suggests that activation in reward regions including vSTR and OFC is most sensitive to the *relative difference* between positive and negative outcomes (10) and between "high-fairness" and "low-fairness" outcomes (7). Therefore, we were most interested in the differential activation to real outcomes that reflected instances when the partner reciprocated compared with those when the partner defected as represented by the contrast Reciprocate > Defect following TRUST decisions. Moreover, we were most interested in *actual* (rather than hypothetical) outcomes, because {Reciprocate} and {Defect} outcomes following KEEP decisions do not represent real gains or real losses, because the participant received 10 MUs *regardless* of partner responses. As such, we would not have expected vSTR and OFC activation between these two fictive outcomes. Thus, first, to measure the brain response to reciprocity, we searched the entire brain for activations to positive vs. negative partner feedback (Reciprocate > Defect, {Reciprocate}>{Defect}) following participant decision to TRUST and to KEEP separately. From this activation map, we selected activated clusters in reward-related regions (e.g., vSTR and OFC) and subjected them to follow-up ROI analysis to clarify the direction and specificity of the effects. Thus, we extracted parameter estimates (β weights, a.u.) from functional ROIs for each individual subject for each of the four outcomes; these β weights were extracted from the functional 10-mm spheres (representing 81 voxels, 648 mm^3) surrounding peak activations in vSTR and OFC for the TRUST/Reciprocate > TRUST/Defect contrast observed in the whole-brain voxelwise search, which were then analyzed with ANOVAs and follow-up *t* tests. The location of the vSTR and OFC activations were confirmed by anatomical atlas from Tzourio-Mazoyer and colleagues and by their consistency with a number of prior fMRI studies showing similar activations in reward (14, 18, 31), reputation (10), and fairness (7). These β weights represent activation averaged across the entire spherical ROI. Second, to examine how vSTR and OFC activation to positive feedback varied as a function of partner type, we used these same functional ROIs and extracted parameter estimates for each individual subject from the contrast of positive and negative feedback (Trust/Reciprocate vs. Trust/Defect) for each of partner type, which were then analyzed with ANOVAs and follow-up *t* tests.

ACKNOWLEDGMENTS. This research was supported by a seed grant from the Brain Research Foundation and National Institutes of Health Grant MH076198.

1. Coleman J (1990) *Foundations of Social Theory* (Harvard Univ Press, Cambridge, MA).
2. Williamson OE (1993) Calculativeness, trust, and economic organization. *J Law Econ* 36:453–486.
3. Cacioppo JT, et al. (2002) Loneliness and health: Potential mechanisms. *Psychosom Med* 64:407–417.
4. Cacioppo JT, Patrick B (2008) *Loneliness: Human Nature and the Need for Social Connection* (W. W. Norton & Company, New York).
5. Fehr E, Camerer CF (2007) Social neuroeconomics: The neural circuitry of social preferences. *Trends Cogn Sci* 11:419–427.
6. Lee D (2008) Game theory and neural basis of social decision making. *Nat Neurosci* 11: 404–409.
7. Tabibnia G, Satpute AB, Lieberman MD (2008) The sunny side of fairness: Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychol Sci* 19:339–347.
8. Krueger F, et al. (2007) Neural correlates of trust. *Proc Natl Acad Sci USA* 104: 20084–20089.
9. Singer T, Kiebel SJ, Winston JS, Dolan RJ, Frith CD (2004) Brain responses to the acquired moral status of faces. *Neuron* 41:653–662.
10. Delgado MR, Frank RH, Phelps EA (2005) Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat Neurosci* 8:1611–1618.
11. King-Casas B, et al. (2005) Getting to know you: Reputation and trust in a two-person economic exchange. *Science* 308:78–83.
12. Rilling J, et al. (2002) A neural basis for social cooperation. *Neuron* 35:395–405.
13. O'Doherty JP (2004) Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Curr Opin Neurobiol* 14:769–776.
14. Knutson B, Adams CM, Fong GW, Hommer D (2001) Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J Neurosci* 21:RC159.
15. Tabibnia G, Lieberman MD (2007) Fairness and cooperation are rewarding: Evidence from social cognitive neuroscience. *Ann N Y Acad Sci* 1118:90–101.
16. Rilling JK, King-Casas B, Sanfey AG (2008) The neurobiology of social decision-making. *Curr Opin Neurobiol* 18:159–165.
17. Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. *Games Econ Behav* 10:122–142.
18. Izuma K, Saito DN, Sadato N (2008) Processing of social and monetary rewards in the human striatum. *Neuron* 58:284–294.
19. Schultz W (2000) Multiple reward signals in the brain. *Nat Rev Neurosci* 1:199–207.
20. McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38:339–346.
21. 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.
22. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275:1593–1599.
23. Berridge KC, Robinson TE (2003) Parsing reward. *Trends Neurosci* 26:507–513.
24. Montague PR, Berns GS (2002) Neural economics and the biological substrates of valuation. *Neuron* 36:265–284.
25. Delgado MR, Li J, Schiller D, Phelps EA (2008) The role of the striatum in aversive learning and aversive prediction errors. *Philos Trans R Soc Lond* 363:3787–3800.
26. Seymour B, Daw N, Dayan P, Singer T, Dolan R (2007) Differential encoding of losses and gains in the human striatum. *J Neurosci* 27:4826–4831.
27. O'Doherty J, Kringelbach ML, Rolls ET, Hornak J, Andrews C (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat Neurosci* 4: 95–102.
28. Kringelbach ML, Rolls ET (2004) The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Prog Neurobiol* 72:341–372.
29. Krueger F, Grafman J, McCabe K (2008) Neural correlates of economic game playing. *Philos Trans R Soc Lond* 363:3859–3874.
30. Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–878.
31. Knutson B, Rick S, Wimmer GE, Prelec D, Loewenstein G (2007) Neural predictors of purchases. *Neuron* 53:147–156.