Behavioral/Cognitive

Reputational Priors Magnify Striatal Responses to Violations of Trust

Elsa Fouragnan,[1] Gabriele Chierchia,[1] Susanne Greiner,[2] Remi Neveu,[3] Paolo Avesani,[2] and Giorgio Coricelli[1,3,4]

[1]Interdepartmental Centre for Mind/Brain Sciences (CIMeC), University of Trento, 38060 Trento, Italy, [2]NeuroInformatics Laboratory (NILab) of Bruno Kessler Foundation, Neuroimaging Laboratory (LNIF) of CIMeC, University of Trento, 38060 Trento, Italy, [3]Centre National de la Recherche Scientifique (CNRS), UMR 5229, University of Lyon, 69003 Lyon, France, and [4]Department of Economics, University of Southern California, Los Angeles, California 90089-0253

Humans learn to trust each other by evaluating the outcomes of repeated interpersonal interactions. However, available prior information on the reputation of traders may alter the way outcomes affect learning. Our functional magnetic resonance imaging study is the first to allow the direct comparison of interaction-based and prior-based learning. Twenty participants played repeated trust games with anonymous counterparts. We manipulated two experimental conditions: whether or not reputational priors were provided, and whether counterparts were generally trustworthy or untrustworthy. When no prior information is available our results are consistent with previous studies in showing that striatal activation patterns correlate with behaviorally estimated reinforcement learning measures. However, our study additionally shows that this correlation is disrupted when reputational priors on counterparts are provided. Indeed participants continue to rely on priors even when experience sheds doubt on their accuracy. Notably, violations of trust from a cooperative counterpart elicited stronger caudate deactivations when priors were available than when they were not. However, tolerance to such violations appeared to be mediated by prior-enhanced connectivity between the caudate nucleus and ventrolateral prefrontal cortex, which anticorrelated with retaliation rates. Moreover, on top of affecting learning mechanisms, priors also clearly oriented initial decisions to trust, reflected in medial prefrontal cortex activity.

Introduction

Trusting others involves risk and uncertainty: people invest a form of good (i.e., money, work, time, etc.) in interactions that can yield a profit or a loss, depending on whether others hold to their end of the bargain (Coleman, 1994). Critically, when others are not contractually committed to doing so, they may be untrustworthy for their own benefit and harm the person that initially placed trust in them (Berg et al., 1995). In financial transactions, investors should then either anticipate this, and not invest money to begin with, or develop efficient strategies to estimate the trustworthiness of others (Camerer and Weigelt, 1988).

Experiments with repeated trust games (RTGs) allow us to empirically observe trust-based dynamics (Chang et al., 2010). Neuroimaging studies using RTGs have shown that, when no prior information on transaction partners is available, the brain's reward circuitry is involved in learning about their type (i.e., their level of trustworthiness), based on the outcomes of previous trust-based interactions (King-Casas et al., 2005). Indeed, reward-related brain regions have been found to respond positively to trustworthiness and negatively to violations of trust (Krueger et al., 2007; Phan et al., 2010; Long et al., 2012). We refer to this as "interaction-based" learning.

However, a second important alternative for investors to efficiently engage in financial decisions is to rely on priors provided by a third party. Such priors may affect the way agents evaluate the outcomes of transactions and thus how they learn about the type of their counterparts. We refer to this as "prior-based" learning. For example, in Web-based transactions, which are increasingly used, investors interact with complete strangers and rely on available reputation priors (e.g., reports on previous transactions, customer reviews, etc.) to predict expected returns and potential risks associated with investments (Kim, 2009). However, while the neural correlates of interaction-based learning to trust have been largely explored, few studies have investigated the neural bases of trust when reputation priors are provided (Delgado et al., 2005; Stanley et al., 2012). No studies to date have directly compared the two forms of trust-based decision making within the same experiment.

To confront this issue, we conducted a functional magnetic resonance imaging (fMRI) experiment in an attempt to characterize the neural activation patterns related to trust-based decisions during RTGs. Two situations were analyzed and compared, one in which we provided information about the social attitude of counterparts (i.e., reputational priors), and one in which no such information was provided. Furthermore, in contrast to a

Received June 29, 2012; revised Dec. 19, 2012; accepted Jan. 7, 2013.

Author contributions: E.F., S.G., P.A., and G. Coricelli designed research; E.F. and S.G. performed research; E.F., S.G., R.N., P.A., and G. Coricelli analyzed data; E.F., G. Chierchia, and G. Coricelli wrote the paper.

This study was supported by the Agence Nationale de la Recherche Française and Provincia Autonoma di Trento. We thank Nadège Bault, Mateus Joffily, and Kévin Vovard for their support in data analysis.

The authors declare no competing financial interests.

Correspondence should be addressed to Elsa Fouragnan, Interdepartmental Centre for Mind/Brain Sciences CIMeC, Via Delle Regole 101, 38060 Mattarello, Trento, Italy. E-mail: elsa.fouragnan@unitn.it.

DOI:10.1523/JNEUROSCI.3086-12.2013

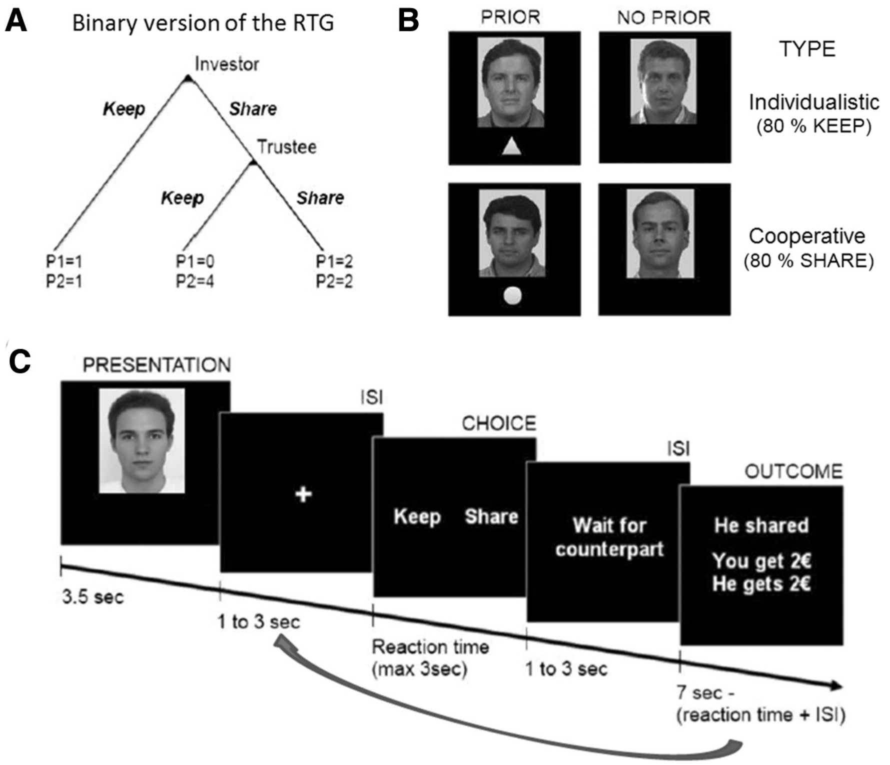

Figure 1. Experimental design. ***A***, One round of the two-player RTG. P1 is the payoff of the participant, who always plays as investor; P2 is the payoff of his counterpart, who plays as trustee. Before each round both players are endowed with €1. The participant moves first and chooses either to "keep" or "share." If he keeps, both players maintain their initial endowments. If he shares the participant's endowment is multiplied by 3 and passed to the counterpart. The trustee then decides whether to share in turn (by returning €2), or to keep (by returning nothing). RTGs consisted of several consecutive rounds with the same counterpart. Participants played with many different counterparts and were told that their counterparts had already made their choices. ***B***, Experimental conditions. Two conditions were adopted: (1) the "type" of counterpart and (2) the presence versus absence of "reputational priors." Types: counterparts could be either "cooperative" or "individualistic" in their (simulated) behavior in RTGs; the former shared and the latter kept in 80% of RTG rounds. Reputational priors: participants were told that cues indicated whether the current counterpart had obtained a high or low score in a social orientation task (triangles indicated low scores, circles indicated high scores). Such priors reliably differentiated between the two counterpart types. ***C***, Time line of the first RTG round. Presentation: face of the counterpart (with a prior or no-prior) was displayed for 3.5 s, and only presented for the first round of an RTG. Fixation: Fixation cross was presented during a jittered ISI. Choice: participants made their choice by pressing "Keep" or "Share." Delay: ISI corresponding to the (simulated) decision of the counterpart. Outcome: outcome of the game and the payoffs of each player.

previous neuroimaging study on the same issue (Delgado et al., 2005), we also manipulated the actual level of trustworthiness demonstrated by counterparts during an RTG, such as to make it consistent with the provided priors. Finally, we used standard fMRI analysis as well as model-free and model-based reinforcement learning (RL) algorithms to approach the problem of social learning and reputation effects. Our main goal was to assess whether and how reputation priors affect RL mechanisms at both the behavioral and neural level.

Materials and Methods

Participants

Twenty male participants (mean age, 29.5 ± 3.53 years) took part in the fMRI experiment; two were removed from the analysis for excessive head movement (see below, fMRI analysis). All of them were healthy; gave written informed consent; had normal or corrected-to-normal vision without any history of psychiatric, neurological, or major medical problems; and were free of psychoactive medications at the time of the study. Participants were told that the experiment aimed at studying decision making in a social context, that they would receive a compensation of €15/h and that the money gained in 10 randomly extracted trials would be added to their compensation. The study was approved by the local institutional ethical committee of the University of Trento.

Task

The experimental task was based on the trust game (TG) (Berg et al., 1995). In one round of our task, each participant played as "investor" with an anonymous counterpart as "trustee." Both players were endowed with €1 before starting a round composed of two stages (Fig. 1*A*): in stage 1 the participant decided whether or not to share his euro with the trustee. If he decided to share, the euro was multiplied by 3 by the experimenter before being allotted to the trustee. In stage 2 the response of the trustee could be to either equally share his money with the investor (1/2 of €4 = €2) or keep his money and return nothing. It follows that if the investor invested and the trustee reciprocated, both players were better off than if the interaction had not occurred at all. However, investing was risky, as if a trustee returned nothing, the investor incurred a loss.

We used a repeated version of this TG (RTG), which consisted in a series of consecutive TG rounds with a same counterpart. However, this alters the nature of the single-shot TG, as RTGs allow for additional strategic maneuvers. For instance, investors tend to invest more (and trustees to reciprocate) in initial rounds of RTGs, than in final rounds or single-shot games (Isaac et al., 1985). For similar reasons, both parties may strategically punish (by not investing) if they believe this might incentivize uncooperative counterparts to review their strategies in future rounds.

Our study intended to minimize the strategic component of trust-related behavior; hence our version of the game differentiated from the typical RTG in a few but important respects. (1) Subjects were informed that trustees had already made their choices, which thus would not have been affected by those of the participant. In other words, participants knew that counterparts were not interactive. This feature should have eradicated any strategic component usually present in RTGs. In reality, the trustees were computer simulations and they reciprocated an investment with fixed probabilities unknown to participants. (2) Another feature was also adopted to make learning independent on participants' actions. In traditional RTGs, when an investor does not trust, the round ends and nothing is learned about the behavior of counterparts. In our study, on the other hand, participants learned about the trustees' choices even when they invested nothing. This adjustment enabled us to keep the amount of feedback fixed (regardless the choice of participants), thus allowing us to compare learning mechanisms between conditions. (3) Finally, to further reduce strategic reasoning, participants did not know how many games composed each RTG with a given trustee but only that RTGs were consecutive and if they were not paired with the same trustee twice in a row, then they would have never encountered the counterpart again. Specifically, we fixed a constant probability of 1/3 to continue the game with a same counterpart; this resulted in a minimum of one and a maximum of eight games with a same trustee.

Then, each trustee was introduced with a picture of his face before a RTG began (Fig. 1*B*). The association between pictures and RTGs was randomized, as was the order of RTGs. To reduce facial information extraction and gender attraction, we assembled a database of colored pictures of 20- to 60-year-old Caucasian men (mean age: 34.05 ± 11.19) controlled for attractiveness, emotion, and racial traits. Then 128 pictures were selected and used with authorization from the FERET database of facial images collected under the FERET program (Phillips et al., 2000). The words "trust" or "trustworthy" were never mentioned during the training session and the experiment.

Experimental conditions

A first key manipulation was that trustees were divided into two predefined types: they could be either "cooperative" or "individualistic." Cooperative trustees would reciprocate 80% of the time, while individualistic counterparts would defect 80% of the time (though participants were not informed of such contingencies). The distinction between types furthermore allowed confronting the cases in which trustees behaved consistently ("Cons") or inconsistently ("Incons") with their types.

The second key feature of our study was whether or not a reputation prior was provided (Fig. 1*B*). In the prior-condition, half of the cooperative and half of the individualistic trustees were flagged, respectively, by a circle and a triangle. These cues signaled their "reputation." Specifically, participants took part in the social valuation orientation (SVO) (Messick and McClintock, 1968; Van Lange et al., 1999) and were told that the distinct cues were based on trustees' scores for the same task. This task distinguishes between different types of SVOs (e.g., cooperative or individualistic). The main difference between each category is the extent to which one cares about his own payoffs and that of the others in social dilemma situations. Finally, for the remaining half of the counterparts, no prior information was provided (no-prior condition).

To ensure no difference in learning scheme in each of the four conditions (Prior/Cooperative, Prior/Individualistic, No-Prior/Cooperative, No-Prior/Individualistic), RTG length and share/keep schedules within each RTG were counterbalanced.

Procedure

Training

Participants received written instructions, took part in a simplified version of the SVO task, and completed a 20 min RTG practice session (20 trials). The experiment was implemented using Presentation software (version 0.70).

Inside the MRI

In the scanner, subjects completed 356 trials (89 for each condition: Prior/Cooperative, Prior/Individualistic, No-Prior/Cooperative, No-Prior/Individualistic), divided in four runs of 20 min. Figure 1*C* shows the time line of the first trial of an RTG. Each RTG started with a 3.5 s display of the face of the trustee (which, only in "prior" conditions, was flagged with a reputational cue). This was followed by a fixation cross and then by a "decision screen," which required participants to choose between one of two options, labeled "share" or "keep." After making their choice, participants waited a jittered interval before an "outcome screen" appeared, displaying the trustee's choice and the corresponding payoffs to both players. For those trials in which participants chose to keep, the outcome screen was still shown.

Analysis

Behavioral data analysis

Behavioral data were analyzed using Stata Statistical Software version 9.2 and the R environment (Development Core Team, 2008). A two-way repeated measure ANOVA was performed to identify differences between conditions for each variable of interest (e.g., decision to trust, payoffs made in each condition). Next, we computed regression analyses using mixed-effects linear models (MEL), in which participants were treated as random effects and hence were allowed to have individually varying intercepts. Parameter estimates (*b*), SE, *t* values and *p* values were reported.

RL models

Model 1: model-free temporal-difference learning. We first used a "model-free" temporal-difference (TD) (model 1) learning algorithm (Rummery and Niranjan, 1994; Sutton and Barto, 1981), which assumes that agents are initially unaffected by the presence of priors, but that, as trials with a counterpart unravel, they may update reward values differently when priors are available as opposed to when they were not available. Participants would sample the reward probability of two choices (Keep or Share) in the Cooperative and Individualistic conditions. We then hypothesized that participants would obtain reliable expectation of these conditions updating the estimated value of each choice with a discounted "step-size." Thus the stochastic prediction error δ, based on the Rescorla–Wagner learning rule (Rescorla and Wagner, 1972) was computed as follows:

$$\delta_t = r_t - Q_{(C,t)} \tag{1}$$

where *r* is the payoff obtained at time *t*, when choosing an option *C* at time *t* or $t + 1$, and *Q* is the value of each choice Share or Keep in each trial. In addition to this, the following learning rule differentially updated the stochastic prediction error in the Prior (*P*) and No-Prior (*NP*) conditions:

$$Q_{(C,t+1)} = Q_{(C,t)} + \alpha^{P} \cdot \delta^{P}{}_{(C,t)} + \alpha^{NP} \cdot \delta^{NP}{}_{(C,t)} \tag{2}$$

The degrees in which δ^P and δ^{NP} influence the new action value are weighted by two learning rates, α^P and α^{NP}, where $0 < \alpha^P, \alpha^{NP} < 1$.

Model 2: model with separate expectations for positive or negative priors. Additionally, we hypothesized that, in the Prior condition, participants may have "optimistic" or "pessimistic" expectations, at the beginning of the game due to the presence of a positive (P^+) or negative Prior (P^-), respectively (Wittmann et al., 2008; Biele et al., 2011) (model 2). Thus, the values of initial choices when playing with a Cooperative or Individualistic counterpart in the prior condition were computed as follows:

$$Q^{P+}{}_{(C,0)} = g^{P+} \cdot \mu\theta_{P+} \cdot N \tag{3}$$

$$Q^{P-}{}_{(C,0)} = g^{P-} \cdot \mu\theta_{P-} \cdot N \tag{4}$$

where g^{P+} g^{P-} are equal to 1 when playing with a counterpart with a positive or negative prior, respectively, and 0 otherwise. θ_{P+} and θ_{P-} are free parameters capturing the optimistic or pessimistic impact of the priors expectation; μ is the expected payoff from choosing randomly among all options, which serves as a normalization constant (in our case $\mu = 1$); and *N* is the number of trials experienced in the learning condition, which is a scaling factor, allowing for the comparison between an expected value decision and the outcome of the decision. On the other hand, in the no-prior condition, only one parameter weighted the initial expected value of choices, $Q^{NP}{}_{(C,0).}$

The Softmax function was then used for the two models to determine the probability of choosing a given choice option given the learned values as follows:

$$\mathrm{p}_1(\mathrm{t}) = \frac{\exp[\mathrm{Q}_1(\mathrm{t})/\beta]}{\exp[\mathrm{Q}_1(\mathrm{t})/\beta] + \exp[\mathrm{Q}_2(\mathrm{t})/\beta]} \tag{5}$$

where β is called a temperature parameter. For high values of β, all actions have almost the same probability (i.e., choices are random), while for low βs the probability of choosing the action with the highest expected reward ($Q_1 > Q_2$) is close to 1.

To generate model-based regressors for the imaging analysis, both learning models were simulated using each subject's actual sequence of rewards and choices to produce per-trial, per-subject estimates of the initial values Q_t and error signals δ_t (Morris et al., 1996; Wittmann et al., 2008). All parameters of interest were implemented in MATLAB R2009 and were estimated using the negative log likelihood of trial-by-trial choice prediction. Model comparisons were performed with the Bayesian Information Criterion, the pseudo r^2 value using the Log likelihood of a random distribution, and tested with the likelihood ratio test.

fMRI method

fMRI data acquisition

A 4 T Bruker MedSpec Biospin MR scanner (CiMEC, Trento, Italy) and an eight-channel birdcage head coil were used to acquire both high-resolution T1-weighted anatomical MRI using a 3D MPRAGE with a resolution of 1 mm^3 voxel and T2*-weighted Echo planar imaging (EPI). The parameters of the acquisition were the following: 34 slices, acquired in ascending interleaved order, the in-plane resolution was 3 mm^3 voxels, the repetition time 2 s, and the echo time was 33 ms. For the main experiment, each participant completed four runs of 608 volumes each. An additional scan was performed in between two different runs to determine the point-spread function that was then used to correct the known distortion in a high-field MR system.

Preprocessing

The first five volumes were discarded from the analyses to allow for stabilization of the MR signal. The data were analyzed with Statistical Parametric Mapping 8 software (SPM8; Welcome Department of Cognitive Neurology, London, UK) implemented in MATLAB R2009 (MathWorks). We used SPM8 for the preprocessing steps. Head motions were corrected using the realignment program of SPM8. Following realignment, the volumes were normalized to the Montreal Neurological Institute (MNI) space using a transformation matrix obtained from the normalization process of the first EPI image of each individual subject to the EPI template. The normalized fMRI data were spatially smoothed with a Gaussian kernel of 8 mm (full-width at half-maximum) in the (x, y, z) axes. Imaging data for participants with head motions exceeding one voxel (3 mm) in transition and 3° in rotation were discarded (Eddy et al., 1996). We also used the xjView package and MRICron to create the pictures presented in the results (version 1.39, Build 4).

fMRI analysis

GLM 1a and 1b. Our first analysis considered the main effect of the presence or absence of reputation priors when a new counterpart is presented for the first time. We used a general linear model (GLM), estimated in three steps: (1) first, an individual blood oxygenation level-dependent (BOLD) signal was modeled by a series of events convolved with a canonical hemodynamic response function. The regressors representing the events of interest were modeled as a boxcar function with onsets at the beginning of each RTG ("Pre") and durations of 3.5 s. For GLM 1a, regressors represented trials in which priors were provided ("Prior_Pre") and no priors were provided ("NoPrior_Pre"). For GLM 1b, regressors represented trials in which priors were provided for a cooperative counterpart ("Prior+_Pre"), priors were provided for individualistic counterparts (Prior-_Pre), and no priors were provided (No-Prior_Pre). For t contrasts, we then computed first-level one-sample t tests comparing trials with and without priors on the basis of the GLM 1a. (2) We then analyzed second-level group contrasts. Our fMRI results were initially thresholded at $p < 0.001$ uncorrected and were subsequently cluster-thresholded at $p < 0.05$, familywise error (FWE). All reported coordinates (x, y, z) are in MNI space. Anatomical localizations were performed by overlaying the resulting maps on a normalized structural image averaged across subjects, and with reference to an anatomical atlas. (3) Finally, we used the MarsBaR toolbox from SPM8 to perform functionally defined (based on the averaged parameter estimates in the cluster found with GLM 1b) region of interest analysis (ROI) and compute percentage signal changes.

GLM 2 model-based fMRI analysis. A second GLM model still focused on the distinction between prior and no-prior conditions but additionally separated between two phases of the RTG: the decision phase and the outcome phase. This allowed us to assess how the impact on the BOLD signal of priors was parametrically modulated by two behaviorally estimated learning measures (from model 2): (1) at time of choice, the parameter Qt, weighted the value of options, on a trial-to-trial basis, depending on RTG history; (2) while δ_t scaled outcomes on the basis of their estimated prediction error. We performed this analysis at the individual level and ran group statistics, taking individual participants as random effects. We then focused on a subset of our resulting brain regions on the basis of effect strength ($p < 0.05$, FWE corrected). Specifically, averaged parameter estimates were extracted from bilateral caudate (MNI coordinates: (−14, 20, 2) and (12, 18, 6)), separating between prior versus no-prior contexts.

GLM 3, violation of trust. In a third GLM we differentiated between consistent (Cons) and inconsistent (Incons) outcomes. We classified consistent outcomes as those rounds in which either (1) participants had kept with individualistic counterparts that defected (Cons−) (distribution of trials: $M = 57 \pm 3$) or (2) they had shared with a cooperative counterpart that reciprocated (Cons+) ($M = 56 \pm 4$ trials); inconsistent outcomes, on the other hand, occurred when either (3) participants had kept with an individualistic counterpart that reciprocated (Incons−) ($M = 14 \pm 4$ trials) or (4) they shared with a cooperative counterpart that defected (Incons+) ($M = 15 \pm 4$ trials), and who thus "violated" their trust.

Functional connectivity analysis. To explore the interplay between the caudate and other brain regions following violations of trust (Incons+), we assessed functional connectivity using psychophysiological analysis (PPI;Friston, 1997; Cohen et al., 2005), which compares the pattern of activity of a seed region to every other regions of the brain. We took the bilateral caudate resulting from the reported GLM 3 (Cons > Incons) as seed regions, as these areas showed highest sensitivity to violations of trust ($t = 6.78$, $p < 0.05$, FWE). Then, we created three regressors: (1) the caudate time course (physiological regressor), (2) an event-related regressor that distinguished between violations of trust in the prior and no-prior conditions (with a boxcar function ranging from the beginning of the outcome phase until the end of the interstimulus interval; ISI), and (3) the interaction term. Additionally, we also conducted a correlation analysis between the retaliation rate for each subject (measured by the percentage of choices to keep after violation of trust when playing with a cooperative partner) and the parameter estimates in left ventrolateral prefrontal cortex (vLPFC) (MNI −40, 42, 4) across subjects. Finally, to examine how striatal responses to violations of trust were related to learning, we plotted individual parameter estimates against the individual learning rates (estimated with model 2 described above).

Results

Behavioral results

Our main goal was to determine whether reputation priors influence initial expectations and decisions in the games, and subsequent learning mechanisms. A repeated measure two-way ANOVA was performed using the type of counterpart (cooperative or individualistic) and prior condition (prior or no-prior) as within participant factors. The percentage of decisions to share was significantly higher with cooperative counterparts ($M = 71.77$, SE ± 4.03) than with individualistic counterparts ($M = 27.34$, SE ± 3.71; $F_{(1,17)} = 174.01$, $p < 0.001$). The results also showed a significant interaction effect of prior with type of counterpart ($F_{(2,35)} = 30.87$, $p < 0.001$). *Post hoc* tests (t tests, Bonferroni corrected) indicated that participants decided to share with cooperative partners more when provided with a prior ($M = 81.09$, SE ± 4.78) than when priors were not provided ($M = 62.45$, SE ± 5.81; $t = 5.89$, $p < 0.001$), whereas they decided to share with individualistic counterparts less in the prior ($M = 18.37$, SE ± 4.66) than in the no-prior condition ($M = 36.3$, SE ± 5.05; $t = 4.23$, $p < 0.002$, Fig. 2*A*). When payoffs are analyzed with type of counterparts and prior condition as within-subject variables, we found that payoffs were significantly higher when playing with cooperative counterparts ($M = 1.43$, SE ± 0.13) than individualistic counterparts ($M = 0.94$, SE ± 0.11; $F_{(1,17)} = 138.32$, $p < 0.001$) and significantly higher in the prior condition ($M = 1.20$, SE ± 0.10) than the no-prior condition ($M = 1.08$, SE ± 0.06; $F_{(1,17)} = 28.98$, $p < 0.001$; Fig. 2*C*).

To examine the effect of the prior condition, trustees' type, the order of the repeated game, and the interactions of such factors on the decision to share (binary-dependent variable), we performed regression analyses using MEL models. The results revealed that participants shared with cooperative counterparts more often compared with individualistic counterparts ($b = 1.29$ (SE ± 0.08), $t = 15.8$, $p < 0.001$), shared less when they did not receive priors ($b = -1.09$ (SE ± 0.09), $t = -12.1$, $p < 0.001$), and shared less over time ($b = -0.12$ (SE ± 0.02), $t = -6.81$, $p < 0.001$). These results suggest that participants took into account reputation priors and played according to the counterpart's level of trustworthiness. Instead, when priors were not available, participants learned counterparts' types on the basis of their actions. Interestingly, we found an interaction effect between trustees' type and the prior condition ($b = 2.27$ (SE ± 0.13), $t = 17.39$, $p < 0.001$). These results indicate that the difference between prior and no-prior conditions was greater when playing with a cooperative than with an individualistic counterpart. Furthermore, even though participants in the no-prior condition adjusted their decisions to their counterparts' type over rounds, they still shared with cooperative counterparts less than when they had priors (Fig. 2*B*). *Post hoc t* test revealed that, in the no-prior condition, in rounds when cooperative counterparts kept, participants subsequently kept more (Mean percentage of decisions to keep = 0.48, SE ± 0.019), whereas they persisted in sharing in the prior condition ($M = 0.2$, SE ± 0.015; $t_{(17)} = -4.99$, $p < 0.001$; Fig. 2*D*). Similarly, when individualistic counterparts shared in a round, participants subsequently shared more when not provided with a prior (Mean percentage of decisions to share = 0.34, SE ± 0.015) than when given a prior ($M = 0.21$, SE ± 0.009; $t_{(17)} = -4.783$, $p < 0.001$; Fig. 2*E*).

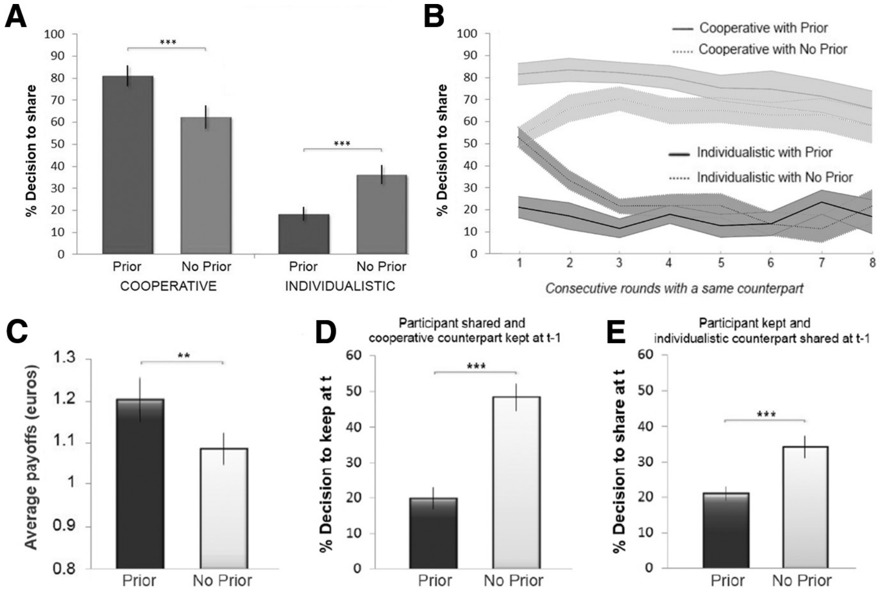

Figure 2. Behavioral results. ***A***, Average percentage of decision to trust across conditions. Mean ± SE of participants' decision to trust (share) are broken down for trustee's type (Cooperative/Individualistic) and prior condition (Prior/No Prior); ***$p < 0.001$. Priors enabled participants to match (on average) their choices with the counterpart's level of trustworthiness. ***B***, Learning dynamics across RTG rounds. Average percentage of the decision to trust for each round when playing with a "cooperative" versus "individualistic" counterpart, and when priors were present versus absent. When participants know nothing of their counterparts they tend to randomize between trusting and not trusting during initial rounds and adjust their choices to their counterparts' type in succeeding rounds. On the other hand, when priors are present, participants tend to rely on them already from early rounds. Shaded areas above and below the curves are SEs. ***C***, Average payoffs in the Prior and No-Prior conditions. Average payoffs ± SE (in €) in Prior/No Prior conditions. When priors are available, participants significantly earn more when they adjust their choices to counterparts' types; **$p < 0.01$. ***D***, Choices following unexpected behavior of cooperative and individualistic counterparts. Average (±SE) of percentage of "keep" choices in prior versus no-prior condition at time *t*, following rounds in which participants shared and a cooperative counterpart violated their trust by deciding to keep (at $t - 1$). Decisions to Keep at time *t* (i.e., retaliation) were less frequent when priors were available. ***E***, Percentage of "share" choices (at *t*) following rounds in which participants had kept and an individualistic counterpart has shared (at $t - 1$).

Results from learning models

A likelihood ratio test revealed that the Prior model (model 2) with separated expectations for cooperative and individualistic counterparts performed better than the classical TD learning model (model 1) ($p < 0.001$) (Additional statistics are reported in Table 1). The best-fitting parameters are shown in Table 2. For these parameters, we found that the average learning rate estimated from trials in the No-Prior condition, α_{NP}, was significantly higher than the average learning rate estimated from trials in the Prior condition α_{P}($t_{(17)} = 2.29$; $p < 0.05$). We also found that the initial value in the Cooperative Prior condition, $Q_{P+}(0)$ was significantly higher than the initial value in the No-Prior condition $Q_{NP}(0)$ ($t_{(17)} = -2.82$; $p < 0.001$), and the initial value in the Individualistic Prior condition, $Q_{P-}(0)$ ($t_{17} = -3.07$; $p < 0.001$). There was no significant difference between the initial value in the Individualistic Prior condition, $Q_{P\text{-}}(0)$, and the initial value in the No-Prior condition $Q_{NP}(0)$ ($t = 0.46$). Finally, we found that the average learning rates estimated for each participant when they kept was higher ($M = 0.46$, SE ± 0.04) than when they shared ($M = 0.38$, SE ± 0.048; $t_{(17)} = -2.27$, $p < 0.05$; Table 2).

fMRI results

Effect of prior at time of counterpart presentation

The contrast (Prior_Pre > NoPrior_Pre) (see Materials and Methods, GLM 1a and 1b) revealed differential activity in the medial PFC (mPFC; 0, 62, 31), to the presence versus absence of any priors when new counterparts were presented ($t = 8.26$; $p < 0.05$, FWE corrected) (Fig. 3*A*; Table 3). Further functional ROI analysis, based on GLM 1b, qualified this activation pattern as responding with increased activity to the presence of priors, regardless of their nature (positive or negative), and decreased activity to their absence (Fig. 3*B*). The opposite contrast (NoPrior_Pre > Prior_Pre) revealed activity in bilateral anterior insula (−36, −4, 15), $t = 3.91$; $p < 0.001$ uncorrected and (38, 3, 10), $t = 3.45$; $p < 0.002$ uncorrected).

Effect of prior at RTG choice

Applying parametric analysis (see Materials and Methods, GLM 2 model-based fMRI analysis) to the functional MRI data, we focused on trial-to-trial weights on decision values as represented by per-trial *Qt* estimate amplitude. We found that decision value estimates were correlated with neural activity in a network consisting of the mPFC (−2, 64, 10) and the dorsolateral prefrontal cortex (dLPFC) (−38, 38, 32), surviving $p < 0.05$, FWE corrected (Fig. 4*A*; Table 3). These two regions reflected the contributions of prior's valence (positive or negative) to the pattern of activity related to the decision to trust (Fig. 4*B*). Moreover, the difference at a neural level between prior and no-prior condition was greater when playing with a cooperative counterpart compared with an individualistic counterpart. This is consistent with the observed behavioral asymmetry of the effect of priors between cooperative and individualistic conditions (Fig. 2*B*).

Table 1. Learning model comparison

Learning model comparison		
	Classical model-free TD learning model	Prior $^+$ and Prior $^-$ expectations RL learning model
BIC	7619	6460
Log likelihood (random model = −4442)	−3809	−3230
Pseudo r^2	0.14	0.273

Bayesian information criterion value (BIC), Log likelihood, and the pseudo r^2 suggest that the Prior $^+$ and Prior $^-$ expectations TD learning model fits the observed behavior better the other TD learning models.

Table 2. Averaged best-fitting parameter estimates (across subjects) SE

Parameter estimate for best behavioral model, depicted as mean ± SE		
	Mean	SE
Learning rate Prior condition α_P	0.3373	± 0.0456
Estimates for Cooperative counterparts	0.327	± 0.0424
Estimates for Individualistic counterparts	0.3475	± 0.0398
Learning rate No Prior condition α_{NP}	0.5075	± 0.0689
Estimates for Cooperative counterparts	0.4686	± 0.0701
Estimates for Individualistic counterparts	0.539	± 0.0599
Estimates learning rates for Invest trials (participants shared)	0.3845	± 0.0459
Estimates learning rates for Non-Invest trials (participants kept)	0.4603	± 0.0476
Softmax inv. Temp Betha β	4.7769	± 0.3149
Initial value Cooperative Prior condition, $Q_{P+}(0)$	1.3814	± 0.1031
Initial value Individualistic Prior condition, $Q_{P-}(0)$	0.9838	± 0.1055
Initial value No Prior condition $Q_{NP}(0)$	1.0641	± 0.126

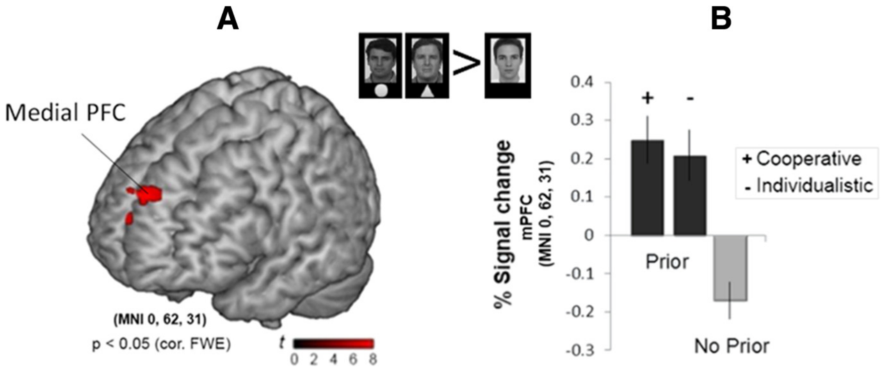

Figure 3. mPFC encodes reputational priors when a new counterpart is first presented. ***A***, Random effect analysis. When contrasting (Prior) > (No Prior) conditions at time of counterpart presentation, activity in the mPFC survived FWE correction, $p < 0.05$. ***B***, Functional ROI analysis in mPFC. Functional ROI analyses further revealed percentage signal changes in the mPFC cortex MNI (0, 62, 31). The figure shows an increased activity when priors were present, regardless of their type, and decreased activity when there were no priors.

Effect of prior at RTG outcome

Across all RTGs, during the outcome phase of the game (see Materials and Methods, GLM 2 model-based fMRI analysis), individually estimated trial-wise prediction errors (positive and negative combined) correlated significantly with BOLD responses in the bilateral caudate in the No-Prior trials only ($p < 0.05$, FWE) (Fig. 5*A*; Table 3). On the other hand, striatal activity appeared to track estimated prediction errors in a more blunted fashion when priors were provided (Fig. 5*A*). Moreover, from a direct comparison between the no-prior and prior conditions, we found higher activity in the left caudate for the no-prior condition compared with the prior condition with group peak MNI coordinates at −12, 20, 8 (Fig. 5*B*).

Pattern of activity related to violation of trust: functional connectivity analysis

Finally, we specified the changes in activity in the caudate related to the effects of violation of trust (e.g., the decisions to keep a cooperative counterpart in response to a decision to trust of a participant) in the prior and no-prior condition (analysis from GLM 3; Table 3). This analysis showed a stronger deactivation of the caudate in the prior condition compared with the no-prior condition ($t = 6.78$; Fig. 6*A*,*C*). However, in contrast with the no-prior condition, striatal deactivations to violation of trust were not reflected in the behavior of our participants. Indeed, the pattern of striatal activity related to violation of trust did correlate with individual learning rates only in the no-prior condition (from the model 2: $r = -0.687$, $p < 0.001$; Fig. 6*D*). No such correlation was found in the Prior condition (Fig. 6*D*).

We used functional connectivity analysis to search (see Materials and Methods, Functional connectivity analysis) for brain areas that could have mediated such striatal responses when reputation priors were provided. We found that left and right vLPFC showed strong functional connectivity with the caudate seed region after violation of trust in the prior compared with no-prior conditions; vLPFC, left (−40, 42, 4), $t = 3.73$; right (38, 46, 4), $t = 6.37$, $p < 0.05$ corrected (Fig. 6*A*; Table 3). Finally, we found that the strength of connectivity between caudate–vLPFC was anticorrelated with participants' decisions to keep following violation of trust (Spearman correlation $r = -0.67$, $p < 0.001$). Moreover, we found that the activity in the vLPFC was inversely correlated with individual retaliation rates (computed as the percentage of Keep over Share choices) after violations of trust ($r = -0.6$, $p < 0.009$; Fig. 6*B*).

Discussion

Reputation-based social decision making has been investigated both by theoretical and empirical studies (Camerer and Weigelt, 1988; Fudenberg et al., 1990; Boero et al., 2009); however, research on its neurocognitive bases is still in its infancy. Though it is rather unlikely that, in daily decisions, people possess absolutely no-prior/contextual information on who they interact with, the growing literature using RTGs in fMRI focused mainly on situations in which strictly no-priors are available (McCabe et al., 2001; King-Casas et al., 2005; Krueger et al., 2007). Only two recent fMRI studies investigated how social priors (i.e., the moral character of their counterparts) affect the way people engage in RTGs (Delgado et al., 2005; Fareri et al., 2012). These studies, however, did not completely isolate the effect of priors on trust (prior-based trust) by confronting them with identical conditions with no priors (interaction-based trust). Our experimental setting is the first to allow this direct comparison. The main goal of our study was to determine whether, and how, reliable reputational priors affect initial decisions and subsequent learning mechanisms at both the behavioral and neural level.

From a behavioral point of view, we show that priors affect decisions to trust in at least two ways: (1) in initial stages of the interaction, participants clearly chose to trust or distrust according to the positive or negative reputation of their counterparts; furthermore (2) players tend to keep relying on

Table 3. Activations correlated with contrasts of interest

						MNI coordinates (mm)		
Analysis/Location	BA	Side	Cluster size	T	p value FWE cor.	X	Y	Z
Prior > No Prior (GLM 1)								
mPFC	10		95	8.26	6.8×10^{-06}	0	62	31
VTA		—	14	3.177	0.0032 unc.	0	−1	−5
No Prior > Prior (GLM 1)								
Anterior insula	44	Left	106	3.912	0.0009 unc.	−36	−4	15
Anterior insula	44	Right	55	3.450	0.0017 unc.	38	3	10
Parametric regression of Choice (GLM 2)								
mPFC	10	—	87	6.562	2.7×10^{-06}	−2	64	10
Lateral PFC	46	Left	122	5.987	7.8×10^{-05}	−38	38	32
Lateral PFC	46	Right	109	6.342	2.1×10^{-06}	30	38	34
Superior parietal lobule	48	Left	43	5.01	6.7×10^{-04}	−38	6	24
Parametric regression at Outcome for the No Prior condition (GLM 2)								
Caudate nucleus	—	Left	77	7.091	8.9×10^{-06}	−14	20	2
Caudate nucleus	—	Right	56	8.298	7.9×10^{-06}	12	16	8
Violation of rust in the Prior condition (GLM 3, Cons > Incons)								
Caudate nucleus	—	Left	82	6.78	2.8×10^{-06}	−10	18	11
Caudate nucleus	—	Right	56	6.34	2.4×10^{-06}	12	21	5

Note: BA, Brodmann area; mPFC, medial prefrontal cortex; VTA, ventral tegmental area.

reputation priors, even when their counterpart's behavior was inconsistent with it. As a consequence, and since priors were accurate predictors of trustworthiness in our study, players earned more when reputational cues were available than when they were not.

mPFC encodes reputational priors

From a neural point of view, our fMRI results revealed that the presentation of a new counterpart yielded enhanced activation in the mPFC when accompanied by a prior (regardless of it signaling a positive or negative reputation). We suggest that the enhanced mPFC activity may reflect the fact the prior information reduced the uncertainty about the behavior of the other faced by participants when beginning a new RTG. Indeed, this region has been previously implicated in uncertainty resolution in interactive contexts (Yoshida and Ishii, 2006). This is furthermore consistent with the inverse activation pattern observed in the insula, which showed stronger activity when priors were not available, consistently with previous findings reporting a role for this region in tracking increased uncertainty (Preuschoff et al., 2008).

mPFC and dLPFC encode the value of reputation priors

At time of choice, the valence of priors elicited dissociable activation patterns when integrated with the behaviorally estimated (from the prior-based RL model) option values (Q_t). Specifically, the mPFC and dLPFC differentially responded to cooperative versus individualistic counterparts, however, only when priors were available.

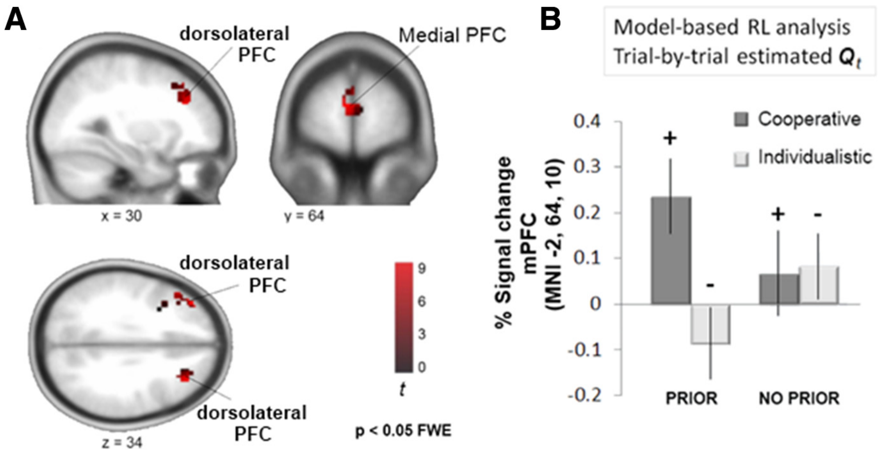

Figure 4. Brain regions parametrically correlated with the estimated "optimistic" and "pessimistic" decision value from the Prior model. ***A***, Random effect fMRI analysis. To look for neural correlates of value signals (Qt) at time of choice, we entered the trial-by-trial estimates of the values of the two stimuli (Share and Keep) into a regression analysis against the fMRI data. We found enhanced activation in mPFC and dLPFC, surviving FWE correction, $p < 0.05$. ***B***, Functional ROI analysis in mPFC. Percentage signal change by condition in the mPFC area represented in ***A***. A similar pattern of activity was found in the dLPFC (not reported). These regions encoded prior valence (positive and negative) that guided decision to trust at time of choice. Error bars indicate SE.

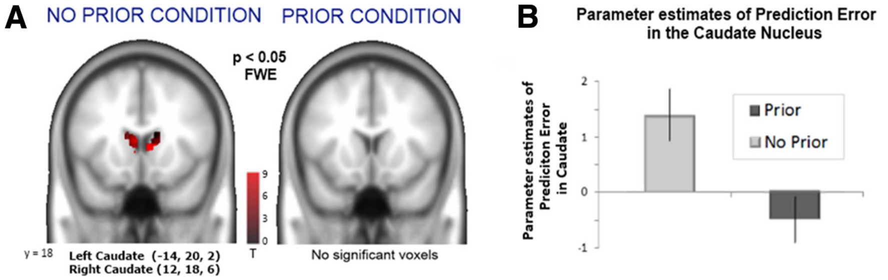

Figure 5. Brain regions parametrically correlated with the estimated prediction error of the best-fitting RL model. ***A***, Random effect fMRI analysis: Activity of the caudate showed significant correlation to the estimated prediction error signal in the no-prior condition ($p < 0.05$ FWE corrected). Such activities were not observed in this brain area in the prior condition. Peak coordinates are given in MNI space. Color bars indicate T values. ***B***, Parameter estimates were extracted from the left caudate (−12, 20, 8) for the direct comparison between prior and no-prior conditions. Caudate activity correlates with prediction error in the no-prior condition only.

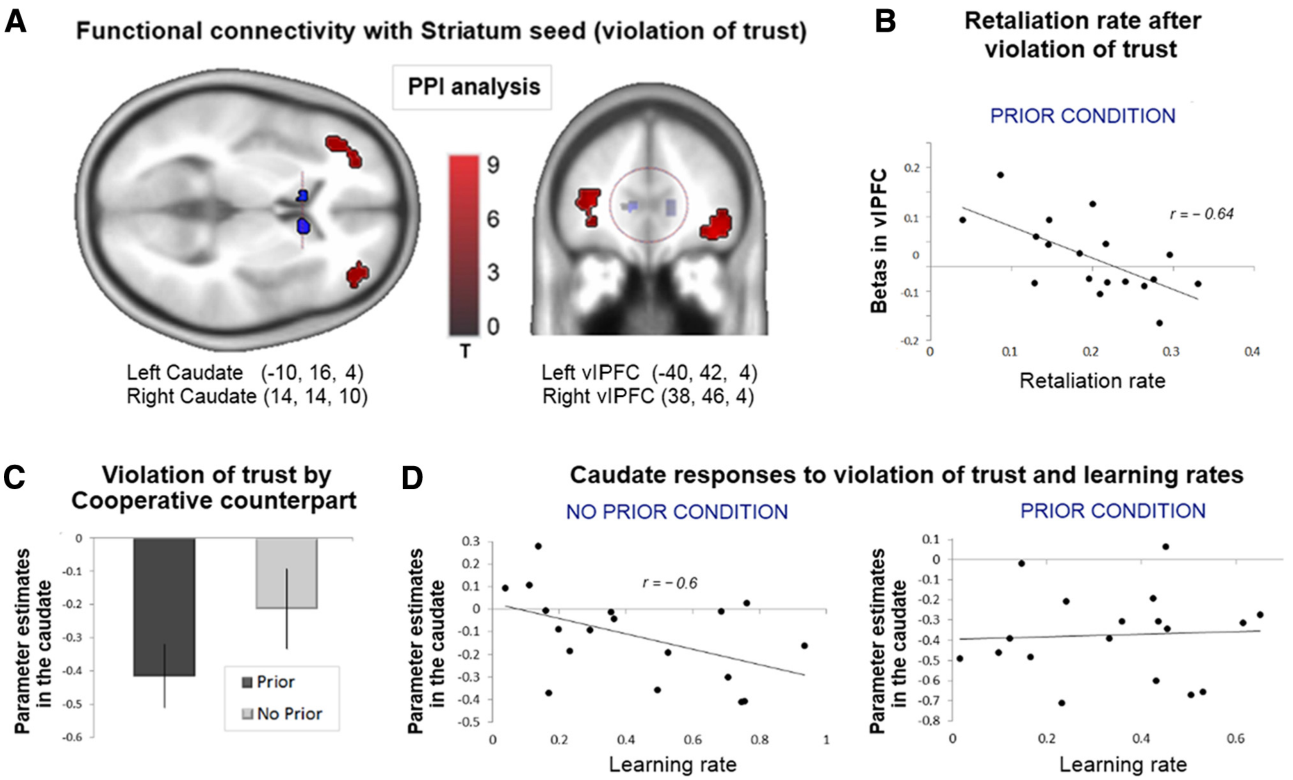

Figure 6. Functional connectivity between the caudate nucleus and vLPFC correlates with the choice to retaliate after violation of trust in the prior condition. ***A***, PPI analysis. With a caudate seed, bilateral vLPFC shows stronger connectivity with this region in the prior compared with the no-prior conditions. ***B***, vLPFC prevents retaliation to violation of trust in the prior condition. vLPFC anticorrelates with retaliation rate in the prior condition after participants experimented violation of trust from a cooperative counterpart. Spearman $r = -0.6$, $p < 0.009$. ***C***, Reputational priors magnify striatal response to violation of trust. The caudate shows a stronger deactivation to violation of trust from a cooperative counterpart in the prior condition compared with the no-prior condition. ***D***, Striatal responses to violation of trust and learning rates. The correlation between caudate and learning rates is significant only in the no-prior condition, thus striatal responses to violation of trust in the prior condition are not reflected in learning.

As reported in previous studies, our results suggest that this brain network keeps track of contextually modulated decision values over trials, and doing so improves participants' performance (Wunderlich et al., 2009).

As reputational priors conveyed information on the social attitudes of counterparts in our study, this activation is also consistent with a well established role of the mPFC in ascribing attitudes to others (Mitchell, 2009), and anticipating their choices (Krueger et al., 2007; Hampton et al., 2008; Coricelli and Nagel, 2009). Thus the mPFC is encoding a first response to reputational priors as well as the effect of priors during subsequent interactions. This is in accordance with findings from humans (Rilling et al., 2002; Hampton et al., 2008) and nonhuman primates (Barraclough et al. 2004) on the role of the PFC in encoding value-related signals in repeated interactions.

Caudate nucleus encodes reward prediction errors only when prior information is not provided

Consistent with previous studies, trial-by-trial prediction errors estimated by RL models correlated with activity in the striatum (McClure et al., 2003; Bunge et al., 2004; O'Doherty et al., 2004; King-Casas et al., 2005; Schönberg et al., 2007), but, critically, only when no priors were available. This confirms a role for the caudate in tracking the difference between expected and obtained outcomes in RTGs, triggering learning. However, when priors were available they appeared to prevent participants from reinforcement-based learning, which was reflected in the reduced covariance between caudate responses and estimated prediction errors.

Priors magnify reward prediction error signals in the caudate nucleus

As regards the striatal activation patterns, these are well aligned with an established role of the striatum in tracking reward contingencies, in both nonsocial (O'Doherty et al., 2004) and social domains (Delgado et al., 2005; King-Casas et al., 2005; Jones et al., 2011). More specifically, the observed patterns are consistent with the idea that the caudate mediates the neural computation of reward prediction error (RPE). Indeed, we observed RPE-pliant signals in the caudate only when no priors were provided, while the same signals appeared blunted when priors were available. Previous studies on nonsocial tasks (Doll et al., 2009, 2011; Li et al., 2011) and social tasks (Delgado et al., 2005; Biele et al., 2011; Fareri et al., 2012) have shown that, when priors are available, participants tended to hinge on to them, and to relatively discount the impact of the outcomes of their past decisions.

However, in addition to the previous studies, our results show that the presence of priors magnifies striatal deactivation to violations of trust (i.e., when a counterpart with positive reputation, as opposed to no reputation, violated trust), rather than blunting their response. Why previous studies did not find such magnified response due to violation of priors requires further investigation, though several hypotheses are possible. For instance, two studies (Delgado et al., 2005; Fareri et al., 2012) focused on the subset of unreliable priors, that is, on priors that carried no information on trustees' actual choices; it is likely that, in such a scenario, participants were gradually learning to disregard such priors, converging toward their extinction rather than exploitation. On the other hand, the opposite may have occurred in a more recent study on the nonsocial domain (Li et al., 2011), in which priors were perhaps too reliable. Indeed, in that study, agents were explicitly instructed on the precise probabilities of outcomes, which may have reduced their surprise when infrequent, though anticipated, losses occurred. In both of these previous studies, the space for learning via priors may have been reduced, as the actual prior-to-reward contingencies appeared either nonexistent (Delgado et al., 2005; Fareri et al., 2012) or already completely exploited (Li et al., 2011). It is also possible that the different methods used to instill priors tapped on different neural mechanisms: Delgado et al. (2005) provided short descriptions of the "moral character" of counterparts, whereas Fareri et al. (2012) used direct evidence from previous experience (i.e., playing a ball task). Such methods of instilling priors may have also made them more salient or intuitive and, as a result, harder to extinguish despite conflicting evidence. On the other hand, our task reported on characteristics of counterparts that were possibly more directly linked to the main task (i.e., the priors were based on results indicating the extent to which one cares about his own payoffs and that of others, SVO task). Further investigation specifically manipulating prior reliability should clarify some of the points of divergence. Until then, the open question in our study regarded the reason as to why striatal deactivations to trust violations were not leading to behavioral adjustments when priors were available.

vLPFC–caudate stronger functional connectivity preventing retaliation

On the other hand, when priors were present, we suggest that the impact on learning of the striatal deactivations to violations of trust may have been disrupted by other brain areas. Our results are in line with attributing this role to the vLPFC, which we found to functionally correlate with such striatal deactivations. In particular, the strength of connectivity between caudate and vLPFC was stronger in the prior compared with the no-prior condition. We thus propose that the vLPFC contributes in maintaining choices aligned with the reliable prior beliefs, when beliefs momentarily conflict with observations. This might occur by compensating for the relatively automatic behavioral changes to RPE signals. In line with this interpretation previous literature has implicated the vLPFC in top-down cognitive control by biasing processing in other brain regions toward contextually appropriate representations (Cohen et al., 1990; Miller et al., 2001). Furthermore, not only the vLPFC plays a role in modulating bottom-up fashion cognition processes, but this area has also been found to play a role in goal-directed behavior (Valentin et al., 2007; Souza et al., 2009).

In conclusion, our study integrates theories and methods from cognitive neuroscience, economics, and reinforcement learning to gain a greater understanding of how reputation priors are encoded in the brain and how they affect learning to trust anonymous others. Our findings suggest that priors influence both initial decisions to trust and the following learning mechanisms involved in repeated interactions. Specifically, the present study showed that reputational priors magnify striatal responses to violations of trust. However, when such priors are reliable, other phylogenetically younger brain regions involved in higher cognition may contribute to keep decisions anchored to those priors, thus relatively discounting the weight of conflicting evidence. The interplay between striatum and ventrolateral prefrontal cortex may prevent unnecessary retaliation when others violate our trust, and thus constitutes an important neurocognitive mechanism that favors social stability.

References

Barraclough DJ, Conroy ML, Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci 7:404–410. Medline

Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. Games Econ Behav 10:122–142. CrossRef

Biele G, Rieskamp J, Krugel LK, Heekeren HR (2011) The neural basis of following advice. PLoS Biol 9:e1001089. CrossRef Medline

Boero R, Bravo G, Castellani M, Squazzoni F (2009) Reputational cues in repeated trust games. J Socio Econ 38:871–877. CrossRef

Bunge SA (2004) How we use rules to select actions: a review of evidence from cognitive neuroscience. Cogn Affect Behav Neurosci 4:564–579. CrossRef Medline

Camerer C, Weigelt K (1988) Experimental tests of a sequential equilibrium reputation model. Econometrica 56:1–36. CrossRef

Chang LJ, Doll BB, van't Wout M, Frank MJ, Sanfey AG (2010) Seeing is believing: trustworthiness as a dynamic belief. Cogn Psychol 61:87–105. CrossRef Medline

Cohen JD, Dunbar K, McClelland JL (1990) On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psychol Rev 97:332–361. CrossRef Medline

Cohen MX, Heller AS, Ranganath C (2005) Functional connectivity with anterior cingulate and orbitofrontal cortices during decision-making. Brain Res Cogn Brain Res 23:61–70. CrossRef Medline

Coleman JS (1994) Foundations of social theory. Boston, MA: Harvard UP.

Coricelli G, Nagel R (2009) Neural correlates of depth of strategic reasoning in medial prefrontal cortex. Proc Natl Acad Sci U S A 106:9163–9168. CrossRef Medline

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. CrossRef Medline

Doll BB, Jacobs WJ, Sanfey AG, Frank MJ (2009) Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. Brain Res 1299:74–94. CrossRef Medline

Doll BB, Hutchison KE, Frank MJ (2011) Dopaminergic genes predict individual differences in susceptibility to confirmation bias. J Neurosci 31: 6188–6198. CrossRef Medline

Eddy WF, Fitzgerald M, Genovese CR, Mockus A, Noll DC (1996) Functional image analysis software- computational olio. In: Proceedings in computational statistics (Prat A, ed), pp 39–49. Heidelberg: Physica-Verlag.

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

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218–229. CrossRef Medline

Fudenberg D, Kreps DM, Maskin ES (1990) Repeated games with long-run and short-run players. Rev Econ Stud 57:555–573. CrossRef

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

Isaac RM, McCue K, Plott C (1985) Public goods provision in an experimental environment. J Pub Econ 26:51–74. CrossRef

Jones RM, Somerville LH, Li J, Ruberry EJ, Libby V, Glover G, Voss HU, Ballon DJ, Casey BJ (2011) Behavioral and neural properties of social reinforcement learning. J Neurosci 31:13039–13045. CrossRef Medline

Kim HH (2009) Market uncertainty and socially embedded reputation. Am J Econ Sociol 68:679–701. CrossRef

King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR (2005) Getting to know you: reputation and trust in a two-person economic exchange. Science 308:78–83. CrossRef Medline

Krueger F, McCabe K, Moll J, Kriegeskorte N, Zahn R, Strenziok M, Heinecke A, Grafman J (2007) Neural correlates of trust. Proc Natl Acad Sci U S A 104:20084–20089. CrossRef Medline

Li J, Delgado MR, Phelps EA (2011) How instructed knowledge modulates the neural systems of reward learning. Proc Natl Acad Sci U S A 108:55–60. CrossRef Medline

Long Y, Jiang X, Zhou X (2012) To believe or not to believe: trust choice modulates brain responses in outcome evaluation. Neuroscience 200: 50–58. CrossRef Medline

McCabe K, Houser D, Ryan L, Smith V, Trouard T (2001) A functional imaging study of cooperation in two-person reciprocal exchange. Proc Natl Acad Sci U S A 98:11832–11835. CrossRef Medline

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

Messick DM, McClintock CG (1968) Motivational basis of choice in experimental games. J Exp Soc Psychol 4:1–25. CrossRef

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202. CrossRef Medline

Mitchell JP (2009) Social psychology as a natural kind. Trends Cogn Sci 13:246–251. CrossRef Medline

Morris JS, Frith CD, Perrett DI, Rowland D, Young AW, Calder AJ, Dolan RJ (1996) A differential neural response in the human amygdala to fearful and happy facial expressions. Nature 383:812–815. CrossRef Medline

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

Phan KL, Sripada CS, Angstadt M, McCabe K (2010) Reputation for reciprocity engages the brain reward center. Proc Natl Acad Sci U S A 107: 13099–13104. CrossRef Medline

Phillips PJ, Moon H, Rizvi SA, Rauss PJ (2000) The FERET evaluation methodology for face-recognition algorithms. IEEE Trans Pattern Anal Mach Intell 22:1090–1104. CrossRef

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

Rescorla RA, Wagner AW (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In: Classical conditioning II: current research and theory (Black AH, Prokasy WF, eds), pp 64–99. New York: Appleton-Century-Crofts.

Rilling J, Gutman D, Zeh T, Pagnoni G, Berns G, Kilts C (2002) A neural basis for social cooperation. Neuron 35:395–405. CrossRef Medline

Rummery GA, Niranjan M (1994) On-line Q-learning using connectionist systems. Technical Report No. 166, University of Cambridge, Engineering Department.

Schönberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci 27:12860–12867. CrossRef Medline

Souza MJ, Donohue SE, Bunge SA (2009) Controlled retrieval and selection of action-relevant knowledge mediated by partially overlapping regions in left ventrolateral prefrontal cortex. Neuroimage 46:299-307. CrossRef Medline

Stanley DA, Sokol-Hessner P, Fareri DS, Perino MT, Delgado MR, Banaji MR, Phelps EA (2012) Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. Philos Trans R Soc B Biol Sci 367:744–753. CrossRef

Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: expectation and prediction. Psychol Rev 88:135–170. CrossRef Medline

Valentin VV, Dickinson A, O'Doherty JP (2007) Determining the neural substrates of goal-directed learning in the human brain. J Neurosci 27: 4019–4026. CrossRef Medline

Van Lange PA (1999) The pursuit of joint outcomes and equality in outcomes: an integrative model of social value orientation. J Personal Soc Psychol 77:337–349. CrossRef

Wittmann BC, Daw ND, Seymour B, Dolan RJ (2008) Striatal activity underlies novelty-based choice in humans. Neuron 58:967–973. CrossRef Medline

Wunderlich K, Rangel A, O'Doherty JP (2009) Neural computations underlying action-based decision making in the human brain. Proc Natl Acad Sci U S A 106:17199–17204. CrossRef Medline

Yoshida W, Ishii S (2006) Resolution of uncertainty in prefrontal cortex. Neuron 50:781–789. CrossRef Medline