Neurocomputational mechanisms of adaptive learning in social exchanges

P. M. Vanyukov[a], **M. N. Hallquist**[b], **M. Delgado**[c], **K. Szanto**[a], **A. Y. Dombrovski**[a,*]

[a]University of Pittsburgh School of Medicine, Department of Psychiatry, Pittsburgh, PA, 15213

[b]Pennsylvania State University, Department of Psychology, State College, PA, 16802

[c]Rutgers University, Department of Psychology, Newark, NJ, 08901

Abstract

Prior work on pro-social and self-serving behavior in human economic exchanges has shown that counterparts' high social reputations bias striatal reward signals and elicit cooperation, even when such cooperation is disadvantageous. This phenomenon suggests that the human striatum is modulated by the other's social value, which is insensitive to the individual's own choices to cooperate or defect. We tested an alternative hypothesis that, when people learn from their interactions with others, they encode prediction error updates with respect to their own policy. Under this policy update account striatal signals would reflect positive prediction errors when the individual's choices correctly anticipated not only the counterpart's cooperation, but also defection. We examined behavior in three samples using reinforcement learning and model-based analyses and performed an fMRI study of striatal learning signals. In order to uncover the dynamics of goal-directed learning, we introduced reversals in the counterpart's behavior and provided counterfactual (would-be) feedback when the individual chose not to engage with the counterpart. Behavioral data and model-derived prediction error maps (in both whole-brain and a priori striatal ROI analyses) supported the policy update model. Thus, as people continually adjust their rate of cooperation based on experience, their behavior and striatal learning signals reveal a self-centered instrumental process corresponding to reciprocal altruism.

Introduction

We often do not know what to expect of others. Although one's social reputation informs these expectations to a degree, we also learn about others from experience. According to formal learning theory, such learning is driven by reward prediction errors – discrepancies between obtained and expected rewards often observed in the striatum and thought to reflect dopaminergic meso-striatal modulation of cortico-striato-thalamic loops (Chase et al., 2015; King-Casas et al., 2005). The effects of rewards and their predictors in social settings are increasingly well-characterized (Bhanji & Delgado, 2014; Bray & O'Doherty, 2007; Kahnt et al., 2008; Klein et al., 2007; Park et al., 2010; Robinson et al., 2013; Schlagenhauf et al., 2012; van den Bos et al., 2009). Even when a desired outcome is not a social reward,

[*]Please direct correspondence to Alexandre Y. Dombrovski, dombax@upmc.edu, 3811 O'Hara St., BT742, Pittsburgh, PA 15213.

however, social predictors still appear to affect goal-directed behavior, leaving open a question of how or whether such information is incorporated into one's learning.

For example, one interesting phenomenon is the effect of social reputation in behavioral and imaging studies of the trust game, where the subject ("investor") invests money with a reciprocating or defecting "trustee". Although the reputation was irrelevant to the trustee's behavior in the cited studies, perceived trustworthiness increased the likelihood of investment and was associated with dampened striatal learning signals (Delgado, Frank, & Phelps, 2005; Fareri, Chang, & Delgado, 2012; *cf.* Fouragnan et al., 2013). Similar effects were shown for social distance (e.g., interacting with a friend versus a stranger: Fareri, Chang, & Delgado, 2015), indicating that the counterpart's perceived social value elicits cooperation regardless of experience. How is the information about these social features incorporated into one's learning? One possibility is that predicting the counterpart's behavior correctly is also rewarding, as it informs one's approach toward others. This form of reinforcement during economic exchanges – referred to here as "policy rewards" - is inherent to cognitive theories of social preference (Camerer, 2003), but has not received much attention in brain research on trust games. Imagine for example that you see an attractively priced used car, but decide against buying it because the dealer seems untrustworthy. You later look up the VIN and learn that the car is a lemon. While your actual reward is zero, you are surely happy that you did not buy the car and will avoid the dealer in future. Hence, to update a policy, one often needs to consider the counterfactual outcomes of one's untaken actions (Boorman, Behrens, & Rushworth, 2011). Consequently, we hypothesized that striatal policy prediction errors during economic exchanges incorporate these counterfactual outcomes.

Although counterfactual representations and learning have been receiving a lot of attention (Boorman et al., 2011; Camille et al., 2004; Coricelli, Dolan, & Sirigu, 2007), these have not been investigated in the context of social decision-making, which by design, i.e., the inclusion of a social counterpart, offers more alternative accounts for which representations should drive learning. In the present study, we sought to test five alternative accounts of how the striatum encodes reinforcement on the trust game and what representation is being updated. (1) The null hypothesis was that instrumental learning only incorporates actual rewards (feedback on the task). (2) We also considered the possibility that the striatum maps the difference between the experienced and best possible outcome, i.e., one possible form of *regret* (Loomes & Sugden, 1982; Lohrenz et al., 2007; Chiu et al., 2008; Nicolle et al., 2011). (3) Furthermore, as people track the counterpart's social value, they may consider *counterpart-oriented counterfactual* choices, i.e., what the outcome would have been had the counterpart made a different decision. (4) Our primary hypothesis, however, was that in addition to learning from actual rewards, people update the value of *their policy* toward a counterpart, incorporating counterfactual outcomes of one's untaken actions. To adjudicate among these four competing accounts, we revealed the would-be decisions of trustees on trials when the subject did not invest. One critical and distinctive prediction consequent to the policy account is that, provided that the subject did not invest, the trustee's defection would be encoded as more rewarding than cooperation. Thus, the direction of the learning signal (or prediction errors) for those trials would be reversed. We also manipulated trustee's cooperation rate to dissect the learning process and adjudicate among alternative models. (5)

Finally, we considered whether subjects track the counterpart's *social value,* as defined by the expected rate of cooperation (Behrens et al., 2008; Fareri, Chang, & Delgado, 2015). This can be thought of as Pavlovian value of the counterpart since it is not conditioned by one's own action history but is still reflected in participant's bias to share or to keep with a particular trustee.

In summary, we investigated the representational underpinnings of striatal signals that mediate adaptive learning in the trust game. We predicted that both behavior and striatal learning signals would be more consistent with updating one's policy than social value- or regret-driven updates.

Methods

Participants

Forty (25 female) middle-aged and older (age: $M = 70.11$, $SD = 10.62$, range 47–95) participants completed the trust task outside (Study 1, $n = 15$) and inside the scanner (Study 3, $n = 25$). To ensure the external validity of our findings across the lifespan, we also replicated our behavioral findings in sample of 29 (19 female) younger participants (Study 2, age: $M = 25.14$, $SD = 6.16$, range 17–45) who performed the task in the scanner (imaging results will be described in a separate paper). The sample sizes were determined on the basis of other neuroimaging paradigms we and others have reported, including that of the trust task. One older participant (Study 3, age > 75) was excluded from the neuroimaging analyses after failing BOLD signal quality checks; specifically, null single subject maps for left versus right responses. Two others were excluded due to excessive motion in the scanner (> *2mm*). All were free from organic brain disease and psychiatric disease, ascertained by an interview, review of medical records and SCID/DSM4. All participants provided written informed consent. The University of Pittsburgh institutional review board approved the studies.

Experimental paradigm

Figure 1A-B illustrates the experimental and single-trial design. Participants completed a modified version of an iterated trust game (Delgado et al., 2005). To enable direct comparisons with previous studies, participants interacted with three fictional trustees described in three separate biographies and, in Studies 1 and 3, with a computer. Each human trustee was introduced with a picture [white, neutral, male faces from the NimStim Face Stimulus Set; (Tottenham et al., 2009)]. A short biography describing trustee's background including one noteworthy event served as an exemplar of trustee's "good", "bad" or "neutral" reputation. In contrast to previous studies (e.g., Delgado et al., 2005), the computer condition was not described as a lottery but as a "computer" player. The computer, therefore, may still have been perceived as having agency vs. a game of chance. In order to enhance experiential learning, we manipulated the rate with which a trustee (or a computer) shared with the participant. All trustees shared 50% of the time for the first 16 trials, as in earlier studies. Then trustees switched to a block of 25% (poor) or 88% (rich) reinforcement schedules, and for the final 16 trials reversed to opposite reinforcement schedules (i.e., rich-to-poor or poor-to-rich). Trial order and the order of "rich" and "poor" reinforcement

schedules across the task were predetermined and four possible orders of the trustees were counterbalanced across the participants.

As before, participants were told that at the beginning of each trial they had $1.00, which they could either keep (and have $1.00 at the end of the trial, regardless of the trustee's decision) or invest with a trustee. If they decided to invest with a trustee, they could receive $1.50 if the trustee shared, or be left with nothing if the trustee decided to keep. Between trustee blocks, participants saw how much money they "won" for that particular block.

A task trial (Figure 1B) started with a fixation cross, followed by the decision phase. During the decision phase, participants viewed a picture of the trustee and two options for keep and share (the location of keep and share options was randomized across trials). In Study 1 (behavior only), the task was self-paced and the decision phase terminated with participant's choice of the option on the left or the right of the screen. In studies 2 and 3 (scanner), participants had up to 2.7 seconds to make their response. After the decision phase, participant's choice was briefly highlighted (300ms) and the outcome was shown. Importantly, in order to examine the mechanisms of experiential learning, we modified the outcome phase from the Delgado et al. study (2005), in which participants received feedback about the trustee's decision only if they chose to share. Instead we showed participants the trustee's decision regardless of their own, so that participants could also learn from counterfactual outcomes and the reinforcement history was kept identical across all participants regardless of their choices. After the keep decisions, the trustee's decision was displayed in grey to indicate that it did not affect the outcome of the trial. The outcome was shown for 1.2 sec. The duration of the next fixation cross was jittered by combining the unused decision time from the previous decision phase and a base duration sampled from an exponential distribution with mean = 810 ms. No durations were allowed to be longer than 1500ms or shorter than 250ms. The trials were blocked by trustee and the computer trials were identical in format to the trustee trials.

Before the experimental session, participants also filled out a seven-point Likert-scale questionnaire, rating the trustworthiness and likability of the counterparts. The same questionnaire and additional assessments were administered at the end of the experiment. E-Prime (Psychology Software Tools, Sharpsburg, PA) was used to present stimulus and control the data acquisition.

Behavioral analyses

There were no missed trials for Study 1 because it was self-paced and too few for Studies 2 and 3 to indicate whether their number differed by block (< 1%). Missed trials were excluded from further analyses. In model-free analyses (not relying on a reinforcement learning model), we examined factors influencing choice in a linear mixed effects framework. A backward model selection procedure was used, re-fitting reduced models and comparing goodness-of-fit statistics. Only the results of models selected by this procedure are reported here. In the first set of analyses, the binary dependent variable was the decision to invest or to keep and the independent variables were trustee type, previous decision made by trustee (representing recent reinforcement history), and the exchange number with trustee. Random effects included a random slope for previous trustee decision (to

accommodate individual differences in learning) and a random intercept for subject (individual differences in cooperation rate). Having determined that the inclusion of an averaged reinforcement schedule (i.e., blocks of 16 trials with varying return rates) did not explain additional variance in the participants' choices in Studies 2 and 3 (see Supplement for details), we analyzed the data pooled across all reinforcement schedules. In the second set of reinforcement model-based analyses, we sought evidence that choices may distinguish between the alternative representational schemes, tested by model comparisons. Specifically, we conceived of the alternatives as different outcome representations (Figure 2, panel A). One hypothesis was that participants track *actual rewards,* i.e., the payoffs are described with respect to how much money participant has by the end of the trial (1). Another alternative was that participants may track a type of *regret* signal when the best possible outcome has not been experienced, i.e., the difference between an actual outcome versus the best possible outcome ($1.50). A third alternative was that participants may track the difference between the outcomes of trustee returning the investment (as the reference point) versus keeping it, i.e., a form of *other-oriented counterfactual* and which in practice only distinguished between the outcomes for when the participant decides to invest the money with the trustee. Finally, we considered the alternative that participants may be tracking the difference between actual and a would-be outcome, resulting from participant's alternative action. For this alternative, the payoffs for the invest decisions were described with respect to how much money participant has by the end of the trial, whereas for the keep decisions, the payoffs were calculated as a counterfactual, i.e., as compared to what would have been the outcome if participant invested. This payoff representation gives rise to the update of the value of one's own *policy,* which in practice distinguished the outcomes for both invest and keep actions of the participant. To reiterate the critical difference between models, all but one of the alternatives predict that if the participant decides to keep on the current trial, the incentive to invest on the next trial should be the same, regardless whether the trustee has kept or shared the money. The policy-oriented representation, however, suggests that following participant's keep decisions, there should be a greater incentive to invest on the next trial, if the trustee shared on the current trial. Thus, we ran additional analyses only on the trials when the participant has made a decision to keep the money and tested whether the likelihood of investing on the next trial was greater for trustee share versus keep decisions. Note that both the regret and trustee-counterfactual models may be considered strong formulations of the competing hypotheses.

Computational modeling

All reinforcement learning (RL) models employed a version of Q-learning with a single hidden state for the expected value of the share action (policy model) or expected reward associated with the trustee (alternative models) on trial t. The value of the keep action was assumed to be updated reciprocally, as shown in previous studies (Dombrovski et al., 2010; Samejima et al., 2005). Q(share) was updated according to the delta rule: $Q(share)_t = Q(share)_{t-1} + \theta \times [reward_t - Q(share)_{t-1}]$, where θ is the learning rate parameter and $reward_t - Q(share)_{t-1}$ is the prediction error. The choice was modeled using a softmax rule, $P(share)_t = 1/(1 + \exp[\ \kappa_s + \kappa_t + \beta \times Q(share)_t]$, where β is a temperature parameter that captures choice stochasticity. We enhanced the choice rule with κ_s, a subject-level parameter, to reflect a bias of the participant to keep or to share regardless of reinforcement and κ_t, a condition-level

parameter, to reflect the participant's bias to keep or to share with a particular trustee, i.e., in this way modeling the effect of social reputation (see Supplement for alternative implementations of trustee- wise bias). Thus, the four models described here have the same four subject-specific parameters (θ, β, κ_s, κ_t) and differ only in their payoff matrices (Figure 2A). A key implication of these differences for imaging analyses and the comparison of the policy model to the rest was that, for the policy model, specifically, relative to keep/return trials (incorrectly predicted defection), the valence of the outcome on keep/keep trials (correctly predicted defection) was positive (Figure 2B), resulting in differential predictions about the trial-by-trial time course of striatal learning signals on keep trials. Thus, while different RL models often predict indistinguishable BOLD time courses when given the same reinforcement (Wilson & Niv, 2015), the counterfactual feedback manipulation enabled us to adjudicate between models based on BOLD data.

We have also evaluated the performance of the Social value (SV) model developed by Fareri and colleagues (2015). The discussion of the SV architecture and the model comparisons are included in the Supplement as secondary analyses.

We used the Variational Bayes Approach (VBA) toolbox implemented in MATLAB (Daunizeau, Adam, & Rigoux, 2014) to test how four reinforcement learning (RL) models fit behavior (likelihood of sharing). We used uninformative Gaussian priors ($M = 0$, $SD = 10$) for all free parameters. We then performed Bayesian model comparisons (BMC) to select the best model (Stephan et al., 2009). An important advantage of the VBA approach is that the relative evidence for different models can be characterized using random effects BMC, while accounting for the full statistical risk incurred [Bayesian omnibus risk, BOR (Rigoux et al., 2014)]. VBA parameterizes the choice history in a state-space framework consisting of dependent variables (i.e., choice time series to be predicted by the model), hidden states (i.e. expected action values tracked over trials), and evolution and observation functions that define the dynamics of hidden state transitions and the model-predicted output, respectively (for details see Daunizeau et al., 2014).

We also simulated data from each model, including SV, to perform parameter recovery and model identifiability analyses. The models' parameters were identifiable (Table S1 and Fig. S1: all Pearson correlations $> .4$, p's $< .001$). We also determined that the models were uniquely distinguishable, i.e., only the model generating a dataset provided the best fit for the dataset (all BORs $< .05$, eps = 1). Details of the simulations are in the Supplement.

fMRI acquisition

BOLD/T2* data were collected with a 3T Siemens Trio Tim scanner with the following parameters: TR = 1670ms, TE = 29ms, FoV = 200 mm, flip=75, 32 3mm slices. We collected 620 volumes across 4 runs.

Image preprocessing

Anatomical scans were registered to the MNI152 template (Fonov et al., 2009) using both affine (FSL FLIRT) and nonlinear (FSL FNIRT) transformations. Functional images were preprocessed using tools from NiPy (Millman & Brett, 2007), AFNI (Cox, 1996), and the FMRIB software library (Smith et al., 2004). First, large transient spikes in voxel time series

were interpolated downward using the AFNI 3dDespike program. Second, slice timing and motion correction were performed simultaneously using a four-dimensional registration algorithm implemented in NiPy (Roche, 2011). Non-brain voxels were removed from functional images by masking voxels with low intensity and by a brain extraction algorithm implemented in FSL BET. The alignment of subjects' functional images to their anatomical scan was computed using the white matter segmentation of each image and a boundary-based registration algorithm (Greve & Fischl, 2009). Functional scans were then resampled into 3mm isocubic voxels and warped into MNI152 template space using the concatenation of the functional-structural and structural-MNI152 transforms. Images were spatially smoothed using a 5-mm full-width at half maximum kernel (FSL SUSAN). A .008 Hz temporal high-pass filter was applied to remove slow-frequency signal changes.

Analyses

Trial-wise prediction error estimates were obtained from VBA posterior. To analyze the effect of prediction errors on the voxelwise blood oxygenation level-dependent (BOLD) signal, their HRF-convolved time-course aligned with feedback was added to choice and feedback regressors in subject-level general linear models in AFNI 3dDeconvolve and 3dREMLfit (Cox, 1996). Importantly, prediction errors from different models were converted to z-scores, in order to avert model-parameter driven scaling differences in the BOLD signal (Lebreton & Palminteri, 2016). Group analyses used the *3dttest*++ program. To control type I error, we thresholded voxelwise tests at $p < .001$ and cluster-thresholded them using a permutation method (*3dttest*++ with the *3dclustsim* option).

Results

Manipulation check: trustworthiness and likeability ratings

Consistent with earlier findings, participants tended to rate good trustees as most trustworthy and bad trustees as the least trustworthy (Figure 3). Detailed analyses are described in the Supplement.

Behavior: effects of reputation

As expected, in all three studies, there was a main effect of trustee (Study 1: $\chi^2 [3] = 18.76$, $p < .001$; Study 2: $\chi^2 [3] = 32.91$; $p < .001$; Study 3: $\chi^2 [3] = 58.13$, $p < .001$). Participants were less likely to invest with the bad trustee compared to the good trustee (Study 1:6 = −0.47, $SE = 0.12$, $p < .001$; Study 2: $b = −0.49$, $SE = 0.09$, $p < .001$; Study 3 : $b = −0.74$, $SE = 0.11$, $p < .001$) with intermediate probabilities of investing with the neutral trustee and computer, except in Study 1, where the likelihood of investing with the computer was lower than but not significantly different from good trustee ($b = −0.081$, $SE = 0.116$, $p = .904$). This is illustrated by Figure 4.

Behavior: adaptive learning, conventional analyses

On average, the trustee returning the investment in the previous round encouraged subsequent investment in all studies (Study 1: $b = 0.57$, $SE = 0.25$, $p = .02$; Study 2: $b = 0.89$, $SE = 0.26$, $p < .001$; Study 3: $b = 0.53$, $SE = 0.20$, $p = .007$). In Studies 2 and 3, there was also a main effect of the number of interactions with a trustee, decreasing the likelihood

of investing with any trustee over time (Study 2: $b = -0.03$, $SE = 0.01$, $p < .001$; Study 3: $b = -0.03$, $SE = 0.01$, $p < .001$). In Studies 2 and 3, there was an interaction of trustee type x number of interactions with trustee (Study 2: χ^2 [1] = 15.66; $p < .001$; Study 3: χ^2 [1] = 31.80; $p < .001$). In Study 2, this interaction indicated that while the likelihood of investment with good trustee decreased over time ($b = -0.81$, $SE = 0.22$, $p < .001$) more dramatically than for bad ($b = 0.53$, $SE = 0.22$, $p = .13$) and neutral trustees ($b = 0.26$, $SE = 0.22$, $p = .47$). In Study 3, this interaction indicated that likelihood of investment with good ($b = -0.80$, $SE = 0.27$, $p = .008$) and computer ($b = -0.59$, $SE = 0.26$, $p = .06$) trustees decreased more dramatically over time than for bad ($b = 0.33$, $SE = 0.26$, $p = .42$) and neutral ($b = -0.09$, $SE = 0.26$, $p = .93$) trustees. Critically, in all three studies, the influence of the trustee's previous action increased with learning (Figure 4B), indicating that participants learned to attend to outcomes rather than reputation over time (Study 1: $b = 0.02$, $SE = 0.01$, $p = .003$; Study 2: $b = 0.03$, $SE = 0.01$, $p < .001$; Study 3: $b = 0.03$, $SE = 0.01$, $p < .001$).

Furthermore, in all three studies, the likelihood of investing on trials following participant's keep decisions increased, if trustee shared versus kept on the current trial, tentatively ruling out all representation alternatives but the policy-oriented: (Study 1: $b = 0.63$, $SE = 0.28$, $p = .024$; Study 2: $b = 0.89$, $SE = 0.27$, $p = .001$; Study 3: $b = 0.63$, $SE = 0.27$, $p = .018$). This incentivizing effect increased with the number of exchanges (trustee's decision × number of exchanges: Study 1: $b = 0.03$, $SE = 0.009$, $p = .001$; Study 2: $b = 0.02$, $SE = 0.008$, $p = .02$; Study 3: $b = 0.03$, $SE = 0.008$, $p < .001$). These latter findings provide evidence of adaptive learning where reinforcement increasingly dominates reputation effects.

Reinforcement learning model comparison

Model comparison of all four alternatives indicated that the policy model dominated the alternatives (Figure 2), although the effect did not withstand the correction for Bayesian omnibus risk (BOR) in the smaller, self-paced Study 1 (Study 1: BOR = 0.386, exceedance probability for policy model (ep) = 0.986; Study 2: BOR = 0.004, ep = 1.0; Study 3: BOR < 0.001, ep = 1.0).

Parametrization of trustee effects in the models

We verified that the means of trustee-level bias parameter (k) from the *policy* model (pooled across all three studies) were significantly different between trustee conditions (χ^2 [2] = 23.44, $p < .001$). The pattern of differences paralleled the effects of trustee conditions on behavior: The bias to invest was smaller for the bad than for the good (t[201] = 4.80, $p < .001$) and for the neutral (t[201] = −2.96, $p = 0.010$) trustees, but not different between good and neutral trustees (t[201] = 1.84, $p = 0.202$). Both the subject-level bias (κ_s) to invest ($b = 0.25$, $SE = 0.03$, $p < .001$) and the subject's rating of the trustee ($b = 0.06$, $SE = 0.01$, $p < .001$) predicted the trustee-level bias. Overall, lower values of trustee-level bias were associated with lower trustworthiness ratings and higher values of trustee-level bias were associated higher trustworthiness ratings (Figure 5).

Neuroimaging

Prediction error (PE) signals—Signed PE signals derived from the policy model were found in the bilateral ventral and dorsal striatum, frontal operculum/posterior insula as well

as a network of paralimbic areas (anterior and mid-cingulate, posterior insula, posterior cingulate/precuneus), and the lateral fronto-parietal network (Figure 6 and Table 1; Supplement includes full maps for PEs derived from alternative models).

Model comparison using neural data—A unique prediction of the policy model is that on keep/keep trials where the subject correctly anticipates the trustee's defection, positive reinforcement (higher signed PEs) is assigned to the taken action compared to keep/return trials where the anticipated defection of the trustee does not occur. PE signals corresponding to the policy PEs were stronger than for regret and trustee-counterfactual models.

Specifically, for an unbiased comparison, we extracted β coefficients from the four maps of model-derived PEs. To avoid circularity, we used the independent mask derived from the meta-analysis of neural correlates of prediction error signal (Chase et al., 2015), consisting of anterior insula, midbrain, dorsal and ventral striatum and visual cortex. To ascertain whether any specific model-estimated PEs indexed stronger activations (i.e., higher β values) we entered model and brain region as first-level predictors in mixed-effects model with subject as a second-level variable with a random intercept (to control for between-subject differences in overall activation strength) and a random intercept for the brain region x subject (to control for between-subject differences in the strengths of activation for the different/nested brain regions). Model type significantly predicted the strength of activation $[\chi^2(3) = 26.24, p < .001)]$. Pairwise contrasts (HSD Tukey adjustment) indicated that policy PEs resulted in greater activation across regions than alternative models (vs. actual rewards: $t(312) = -3.80, p = 0.001$; regret: $t(312) = 4.58, p < 0.001$; trustee: $t(312) = 4.01, p < 0.001$). That is, the signal derived from this model appeared to best capture brain activity than the signal derived from other models. We did not test the SV model against neural data, given that its key SV parameter was not significantly different from 0 and could not affect PE signals. To ascertain that these findings were not an artifact of our modeling choices, we also verified that maps for the non-model-based contrast corresponding to the policy update hypothesis (congruent [invest-return and keep-keep] vs. incongruent [keep-return and invest-keep] trials) were similar, but statistically weaker (striatum: $t_{max} = 4.52, k = 18$ at $p_{voxelwise} < 10^{-4}$). Furthermore, in model-free analyses, we evaluated the activation differences with a contrast between the four types of trial outcomes: subject invests/trustee returns, subject invests/trustee keeps, subject keeps/trustee returns, subject keeps/trustee keeps. From the group maps generated by *3dMEMA*, we extracted the β coefficients using an unbiased ventral striatum mask from a PE meta-analysis (Chase et al., 2015) and the meta-analytic estimates of variance (τ^2). Illustrated by Figure 7, comparisons indicated that striatal activation was greater bilaterally when subjects invested and trustee returned the investment (vs. subject invested/trustee kept: *z*ieft striatum = 20.24; *z*right striatum = 25.56; vs. subject kept/trustee returned: *z*ieft striatum = 29.47; *z*right striatum = 18.61; all $ps < .001$) and when subjects kept and trustee kept the investment (vs. subject invested/trustee kept: *z*ieft striatum 13.21; *z*right striatum 13.24; vs. subject kept/trustee returned: *z*ieft striatum 16.86; *z*right striatum = 12.30; all $ps < .001$). There was slightly greater activation in the left striatum for mutual cooperation vs. defection outcomes ($z = 7.21, p < .001$), but no difference in activation for the right striatum, or for the two incongruent outcomes. Thus, the critical

observation here is that striatal responses to correctly predicted trustee defections are positive.

Effects of Trustee type—Because participants were less likely to invest with the bad trustee vs. all others, we also examine this contrast and its interaction with prediction errors. We found no evidence of trustee effects or of their interaction with prediction errors (see Supplement for details). Thus, reputation did not appear to directly interfere with instrumental learning signals.

Discussion

Our findings indicate that adaptive learning during social exchanges depends on tracking the success of one's approach (policy) toward the counterpart and is unlikely to be driven merely by actual rewards or updates of the counterpart's Pavlovian "social value". This conclusion was supported by model-free behavioral analyses and reinforcement learning model comparison in three samples and by analyses of neural activation. Compared to alternative models, learning signals derived from the policy model, which tracked the success of a given approach, indexed stronger activation of the striatum and other nodes in the canonical prediction error network, e.g., anterior insula, caudate, and thalamus. Critically, when participants correctly predicted the counterparts' defection, we observed positive striatal prediction errors. Our study also confirmed the presence of, in this case, irrelevant reputation effects (Delgado et al., 2005; Fareri et al., 2012). However, our data suggested that reputation effects do not interact with adaptive instrumental learning, but rather exert an additive Pavlovian influence on cooperation rates.

Persistent effects of social reputation on participants' decisions in the trust task have led earlier studies to focus on whether participants estimate the social value or pro-sociality of the counterpart, indicated by the trustee's return on the investment. In these studies, participants also learned of trustee's decisions only from their decisions to invest (Delgado et al., 2005; Fareri et al., 2012, 2015). Thus, both the experimental design and the modeling approach implied a Pavlovian account of social decision-making in this task, i.e., learning about the counterpart as a stimulus, without regard to one's own actions. In the present study, by showing participants outcomes for both actions and introducing reversals into the trustees' behavior, we were able to investigate the extent to which decision-making was accounted for by an instrumental process, as well as to examine representational schemes that incorporated counterfactual thinking.

Importantly, our findings indicate that the behavior and learning signals in the task are best described by a policy that anticipates the trustees' actions. This interpretation is supported by our neuroimaging findings: prediction errors derived from a model tracking the successes of its own decisions resulted in more robust maps than those derived from the other three representational schemes. Critically, model-free analyses of neural signals associated with the four different trial outcomes indicated stronger positive striatal responses when participant's decisions matched that of the trustee, including strongly positive responses to correctly predicted defections. Taken together, these results illuminate the reference frame for striatal instrumental learning signals during social exchanges: they represent updates on

the individual's action policy, with negative responses to unpredicted cooperation and positive responses to predicted defection. Effects of trustee reputation on behavior, captured by a trustee-specific bias parameter provide support for independent, possibly parallel, processes: an instrumental learning process as revealed by our design, and Pavlovian approach/avoidance toward the counterpart influenced by reputation.

Interestingly, in the more recent examination of prediction errors in the trust task (Fareri et al., 2015), where social distance from the counterpart was controlled, social value (SV) model-derived prediction error maps were similar to those observed in this study, even if less robust statistically. The computational model in Fareri et al. tracked participant's beliefs about the probability of the trustee sharing and, thus, prediction errors reflected the difference between the outcome and that expectancy. This is in contrast to the policy model, which posited that participants track the expected value of the action to invest. However, similarly to the models presented here, Fareri et al. model included a social value parameter, which enhanced reciprocation with high-reputation counterparts, affecting choice but not learning. In terms of learning, the predictions of the policy model and Fareri's social value model are nearly identical on trials where the participant shares, but on keep/return trials the policy model predicts a negative outcome, while the social value model does not. This difference must be responsible for the fact that, when confronted with counterfactual feedback on keep trials, the social value model could not provide a precise account of participants' behavior.

It may be important acknowledge that our study, in which one learns about would-be or counterfactual outcomes, reflects only a subset of social situations in which one learns second-hand about the trustworthiness of others based on a transaction in which one was not directly involved. We believe that this learning may be critical to those of individual's interactions that can be framed as investment decisions of time and/or other resources with others (as well as, more conventionally thought of as investments in the financial world). In the introduction, we used an example of a one-shot counterfactual learning: a car, which one decided not to purchase, has turned out to be a lemon and, therefore, was a good decision. We would argue that one's decisions in a social setting, often characterized by multiple transactions, can be informed in a similar manner. For example, more specific to the academic setting, a colleague may solicit collaboration on an article (or a grant) which may require a significant investment of time and effort. Having refused this offer, one can learn about the article's publication (or a grant's acceptance) that may, in turn, influence one's future decisions on collaboration.

More specifically, our findings demonstrate that adaptive learning in these situations may depend critically on credit assignment, the process by which reward value is attributed to one's decisions or stimuli, extensively studied in non-social animal learning (Boorman et al., 2011; Boorman et al., 2009; Daw et al., 2011; Noonan et al., 2010; Walton et al., 2010). Correct credit assignment depends on the ability to set up the state space, or a map of available actions and possible outcomes in a given environment. The state space is thus organized around one's policy, or an algorithm by which one selects which action to take, which in turn controls outcomes (state transitions). In the social setting, we would argue that the counterpart, e.g., a trustee, is viewed as an environment where one forages. The

counterpart's social characteristics induce a Pavlovian stimulus-outcome belief eliciting approach or avoidance. We examined striatal prediction error (PE) signals - the best-validated neural index of reward learning - to adjudicate among alternative representational schemes. The ability to find the best strategy and overcome irrelevant social biases, however, is likely to depend on multiplexing of information about own and others' choices and outcomes in the dorsal anterior cingulate cortex [ACCs/ACCg (Apps, Rushworth, & Chang, 2016)], also indexed by PEs in our study. For example, Haroush and colleagues, recording form the monkey ACCs, described two cell populations whose firing rates predict not only the monkey's own choice, but also the counterpart's choice in the prisoner's dilemma (Haroush & Williams, 2015).

One limitation of this study was the lack of a live trustee, which precluded us from examining how participants represented the impact of their own actions on the counterpart. However, reducing such strategic behavior in participants was necessary to closely examine trial-by-trial learning. Because the trials for each trustee had to be blocked rather than interleaved to enhance learning, we had to sacrifice the power to examine neural correlates of reputation and its putative interactions with PE signals. It is also important to note that, detracting from our ability to capture the trustee effects, trustee was a block variable which varied by run. Therefore unique baselines for trustees could not be recovered, in contrast to studies with a mixed trial design. We also did not query participants whether they were consciously aware of the changes in reinforcement schedules. The results of our computational models, however, suggest that their behavior reflected learning of the different reinforcement rates. Furthermore, participants' exposure to counterfactual outcomes varied with their cooperation rate, raising questions about the robustness of our results vis-à-vis different strategies. This issue is partly mitigated by including the random effect of person in both model-free analyses and RL model comparisons. Finally, although age-related differences have been observed in other economic games (e.g., Brown & Ridderinkhof, 2009), we have not systematically investigated these differences in this paper and instead report the replication of the behavioral findings in the older participants from studies 1 and 3, and also in the younger sample in Study 2.

In summary, we found that as humans learn whether to cooperate with others from experience, their striatal learning signals reflect reinforcement by the correctness of their predictions about what others will do, a process that may underlie reciprocal altruism.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

Apps MA, Rushworth MF, & Chang SW (2016). The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. Neuron, 90(4), 692–707. [PubMed: 27196973]

Behrens TEJ, Hunt LT, Woolrich MW, & Rushworth MFS (2008). Associative learning of social value. Nature, 456(7219), 245–249. 10.1038/nature07538 [PubMed: 19005555]

Bhanji JP, & Delgado MR (2014). The social brain and reward: social information processing in the human striatum. Wiley Interdisciplinary Reviews: Cognitive Science, 5(1), 61–73. [PubMed: 24436728]

Boorman ED, Behrens TE, & Rushworth MF (2011). Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. PLoS Biol, 9(6), e1001093. [PubMed: 21738446]

Boorman ED, Behrens TE, Woolrich MW, & Rushworth MF (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron, 62(5), 733–743. [PubMed: 19524531]

Bray S, & O'Doherty J (2007). Neural coding of reward-prediction error signals during classical conditioning with attractive faces. J Neurophysiol, 97(4), 3036–3045. 10.1152/jn.01211.2006 [PubMed: 17303809]

Brown SB, & Ridderinkhof KR (2009). Aging and the neuroeconomics of decision making: A review. Cognitive, Affective, & Behavioral Neuroscience, 9(4), 365–379.

Camerer CF (2003). Behavioural studies of strategic thinking in games. Trends in Cognitive Sciences, 7(5), 225–231. 10.1016/S1364-6613(03)00094-9 [PubMed: 12757825]

Camille N, Coricelli G, Sallet J, Pradat-Diehl P, Duhamel J-R, & Sirigu A (2004). The Involvement of the Orbitofrontal Cortex in the Experience of Regret. Science, 304(5674), 1167–1170. 10.1126/science.1094550 [PubMed: 15155951]

Chase HW, Kumar P, Eickhoff SB, & Dombrovski AY (2015). Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. Cognitive Affective and Behavioral Neuroscience. 10.3758/s13415-015-0338-7

Chiu PH, Lohrenz TM, & Montague PR (2008). Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. Nature Neuroscience, 11(4), 514–520. [PubMed: 18311134]

Coricelli G, Dolan RJ, & Sirigu A (2007). Brain, emotion and decision making: the paradigmatic example of regret. Trends in Cognitive Sciences, 11(6), 258–265. 10.1016/jztics.2007.04.003 [PubMed: 17475537]

Cox RW (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computational Biomedical Research, 29(3), 162–173.

Daunizeau J, Adam V, & Rigoux L (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. PLoS Comput Biol, 10(1), e1003441 10.1371/journal.pcbi.1003441 [PubMed: 24465198]

Daw ND, Gershman SJ, Seymour B, Dayan P, & Dolan RJ (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron, 69(6), 1204–1215. [PubMed: 21435563]

Delgado MR, Frank RH, & Phelps EA (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. Nature Neuroscience, 5(11), 1611–1618.

Dombrovski AY, Clark L, Siegle GJ, Butters MA, Ichikawa N, Sahakian BJ, & Szanto K (2010). Reward/Punishment reversal learning in older suicide attempters. American Journal of Psychiatry, 167(6), 699–707. 10.1176/appi.ajp.2009.09030407 [PubMed: 20231320]

Fareri DS, Chang LJ, & Delgado MR (2012). Effects of direct social experience on trust decisions and neural reward circuitry. Frontiers in Neuroscience, 6, 1–17. [PubMed: 22294978]

Fareri DS, Chang LJ, & Delgado MR (2015). Computational substrates of social value in interpersonal collaboration. The Journal of Neuroscience, 35(21), 8170–8180. 10.1523/JNEUROSCI.4775-14.2015 [PubMed: 26019333]

Fonov VS, Evans AC, McKinstry RC, Almli CR, & Collins DL (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage, (47), S102.

Fouragnan E, Chierchia G, Greiner S, Neveau R, Avesani P & Coricelli G (2013). Reputational priors magnify striatal responses to violations of trust. The Journal of Neuroscience, 33(8), 3602–3611. [PubMed: 23426687]

Greve DN, & Fischl B (2009). Accurate and robust brain image alignment using boundary-based registration. Neuroimage, 48(1), 63–72. [PubMed: 19573611]

Haroush K, & Williams ZM (2015). Neuronal prediction of opponent's behavior during cooperative social interchange in primates. Cell, 160(6), 1233–1245. 10.1016/_j.cell.2015.01.045 [PubMed: 25728667]

Kahnt T, Park SQ, Cohen MX, Beck A, Heinz A, & Wrase J (2008). Dorsal Striatal-midbrain Connectivity in Humans Predicts How Reinforcements Are Used to Guide Decisions. Journal of Cognitive Neuroscience, 21(7), 1332–1345. 10.1162/jocn.2009.21092

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(5718), 78–83. 10.1126/science.1108062 [PubMed: 15802598]

Klein TA, Neumann J, Reuter M, Hennig J, von Cramon DY, & Ullsperger M (2007). Genetically determined differences in learning from errors. Science, 318(5856), 1642–1645. 10.1126/science.1145044 [PubMed: 18063800]

Lebreton M, & Palminteri S (2016). Revisiting the assessment of inter-individual differences in fMRI activations-behavior relationships. BioRxiv, 036772.

Lohrenz T, McCabe K, Camerer CF, & Montague PR (2007). Neural signature of fictive learning signals in a sequential investment task. Proceedings of the National Academy of Sciences, 104(22), 9493–9498.

Loomes G, & Sugden R (1982). Regret theory: An alternative theory of rational choice under uncertainty. The Economic Journal, 92(368), 805–824.

Millman KJ, & Brett M (2007). Analysis of functional magnetic resonance imaging in Python. Computing in Science & Engineering, 9(3), 52–55.

Nicolle A, Bach DR, Driver J, & Dolan RJ (2011). A role for the striatum in regret- related choice repetition. Journal of Cognitive Neuroscience, 23(4), 845–856. [PubMed: 20433245]

Noonan MP, Walton ME, Behrens TE, Sallet J, Buckley MJ, & Rushworth MF (2010). Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. Proc Natl Acad Sci U S A, 107(47), 20547–20552. 10.1073/pnas.1012246107 [PubMed: 21059901]

Park SQ, Kahnt T, Beck A, Cohen MX, Dolan RJ, Wrase J, & Heinz A (2010). Prefrontal Cortex Fails to Learn from Reward Prediction Errors in Alcohol Dependence. Journal of Neuroscience, 30(22), 7749–7753. 10.1523/JNEUROSCI.5587-09.2010 [PubMed: 20519550]

Psychology Software Tools, Inc. [E-Prime 3.0]. (2016). Retrieved from http://www.pstnet.com.

Rigoux L, Stephan KE, Friston KJ, & Daunizeau J (2014). Bayesian model selection for group studies — Revisited. NeuroImage, 84, 971–985. 10.1016/_j.neuroimage.2013.08.065 [PubMed: 24018303]

Robinson OJ, Overstreet C, Charney DR, Vytal K, & Grillon C (2013). Stress increases aversive prediction error signal in the ventral striatum. Proceedings of the National Academy of Sciences, 110(10), 4129–4133. 10.1073/pnas.1213923110

Roche A (2011). A four-dimensional registration algorithm with application to joint correction of motion and slice timing in fMRI. IEEE transactions on medical imaging, 30(8), 1546–1554. [PubMed: 21427017]

Samejima K, Ueda Y, Doya K, & Kimura M (2005). Representation of action-specific reward values in the striatum. Science, 310(5752), 1337–1340. 10.1126/science.1115270 [PubMed: 16311337]

Schlagenhauf F, Rapp MA, Huys QJ, Beck A, Wustenberg T, Deserno L, … Heinz A (2012). Ventral striatal prediction error signaling is associated with dopamine synthesis capacity and fluid intelligence. Human Brain Mapping. 10.1002/hbm.22000

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, … & Niazy RK (2004). Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage, 23, S208–S219. [PubMed: 15501092]

Stephan KE, Penny WD, Daunizeau J, Moran RJ, & Friston KJ (2009). Bayesian model selection for group studies. NeuroImage, 46(4), 1004–1017. https://doi.org/10.1016Zj.neuroimage.2009.03.025 [PubMed: 19306932]

Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, … Nelson C (2009). The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Research, 168(3), 242–249. [PubMed: 19564050]

van den Bos W, van Dijk E, Westenberg M, Rombouts SARB, & Crone EA (2009). What motivates repayment? Neural correlates of reciprocity in the Trust Game. Social Cognitive and Affective Neuroscience, 4(3), 294–304. 10.1093/scan/nsp009 [PubMed: 19304843]

Walton ME, Behrens TE, Buckley MJ, Rudebeck PH, & Rushworth MF (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron, 65(6), 927–939. [PubMed: 20346766]

Wilson RC, & Niv Y (2015). Is model fitting necessary for model-based fMRI? PLOS Computational Biology, 11(6), e1004237 10.1371/journal.pcbi.1004237 [PubMed: 26086934]
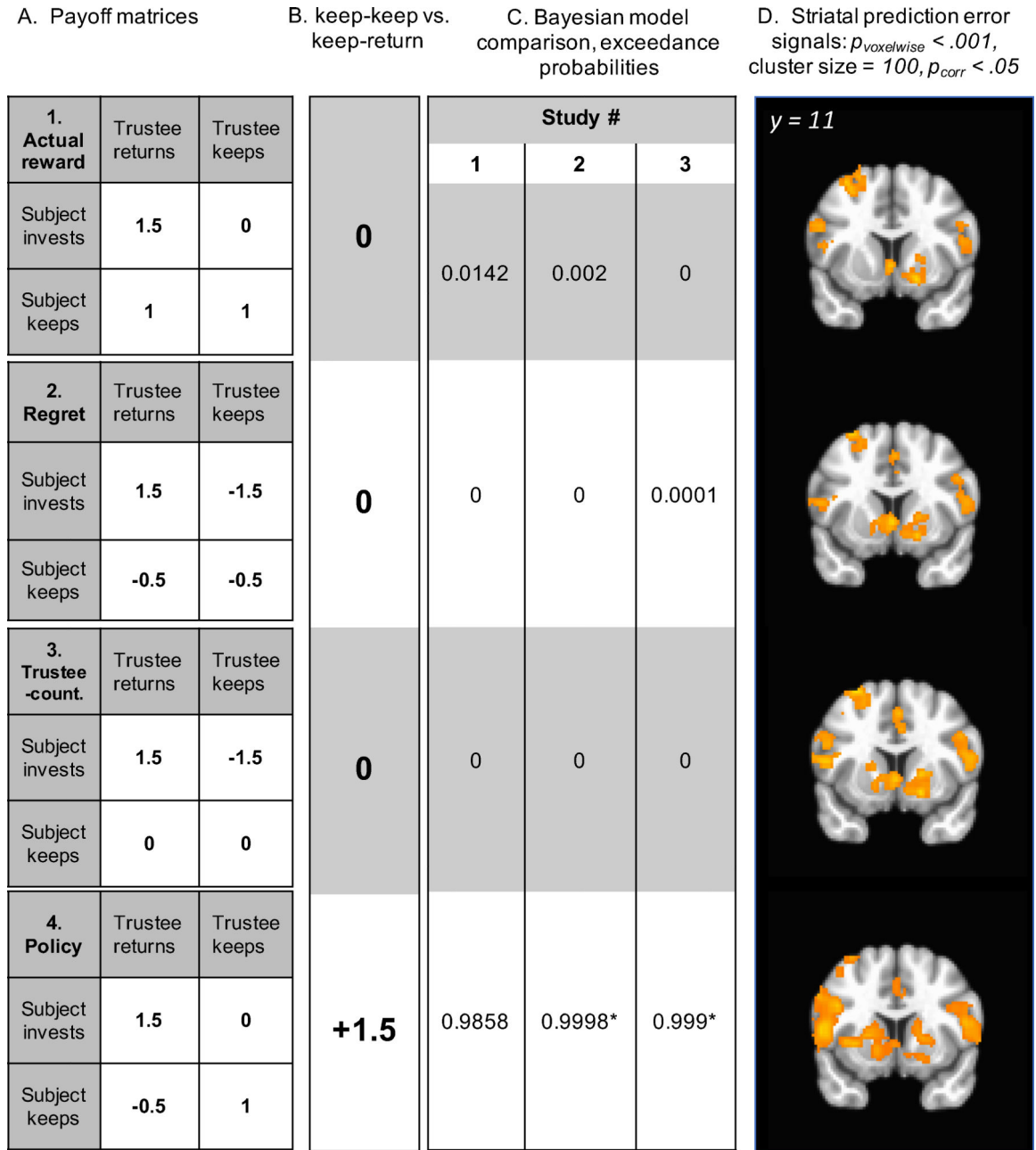
A.

Trustee Type: Presentation blocked and order counterbalanced across participants	Reinforcement Schedule: Order constant across participants			
	Block 1 (16 trials)	Block 2 (16 trials)	Block 3 (16 trials)	192 total trials
good	50%	88% (rich)	25% (poor)	48 trials
bad	50%	25% (poor)	88% (rich)	48 trials
neutral	50%	25% (poor)	88% (rich)	48 trials
computer	50%	88% (rich)	25% (poor)	48 trials

B.
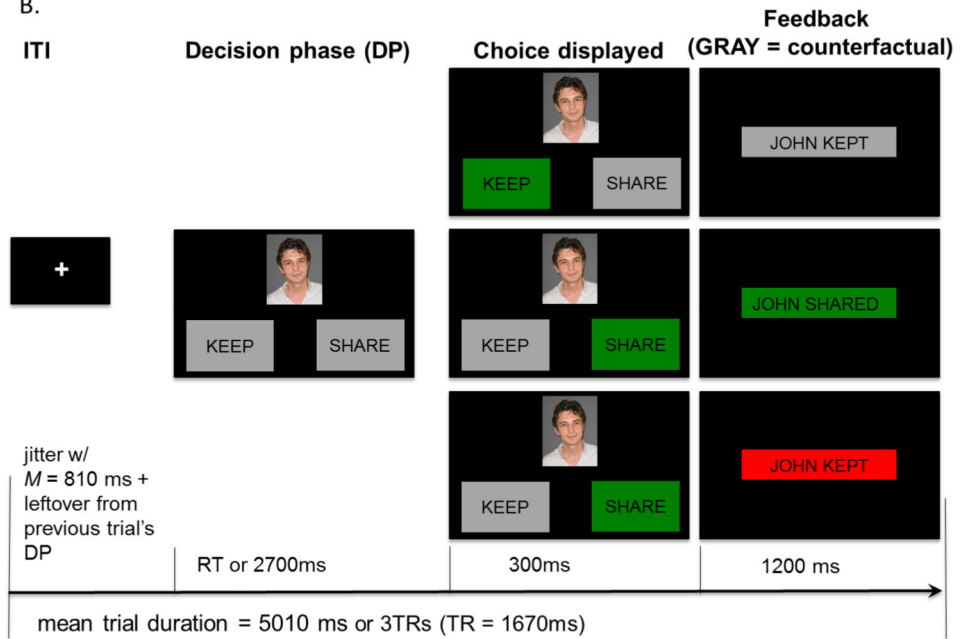
Figure 1.
(A) Reinforcement schedule by trustee and block; (B) Trial design. Counterfactual feedback is presented to participants after "KEEP" decisions (top panel).

A. Payoff matrices **B. keep-keep vs. keep-return** **C. Bayesian model comparison, exceedance probabilities** **D. Striatal prediction error signals: $p_{voxelwise} < .001$, cluster size = 100, $p_{corr} < .05$**

1. Actual reward	Trustee returns	Trustee keeps
Subject invests	1.5	0
Subject keeps	1	1

2. Regret	Trustee returns	Trustee keeps
Subject invests	1.5	-1.5
Subject keeps	-0.5	-0.5

3. Trustee-count.	Trustee returns	Trustee keeps
Subject invests	1.5	-1.5
Subject keeps	0	0

4. Policy	Trustee returns	Trustee keeps
Subject invests	1.5	0
Subject keeps	-0.5	1

B. keep-keep vs. keep-return:
- 1. Actual reward: 0
- 2. Regret: 0
- 3. Trustee-count.: 0
- 4. Policy: +1.5

	Study #		
	1	2	3
1	0.0142	0.002	0
2	0	0	0.0001
3	0	0	0
4	0.9858	0.9998*	0.999*

y = 11

Figure 2.
(A) Payoff matrices by model; (B) difference between correctly anticipated (keep/keep) and incorrectly anticipated trustee defection (keep/return) shaping predicted striatal learning signal time course on those trials, (C) exceedance probabilities from the Bayesian model comparison, (D) neural responses in ventral striatum to model-derived prediction errors (PE) ($p_{VOXeiwise} < 001$, cluster size = 100 voxels, $p_{corr} < .05$). 1) Actual-rewards model: actual rewards in the task; 2) Regret model: difference between actual rewards and the maximum outcome ($1.5); 3) Trustee-counterfactual model: difference between actual rewards and would-be outcome from trustee's alternative action; 4) Policy model (**best model**):
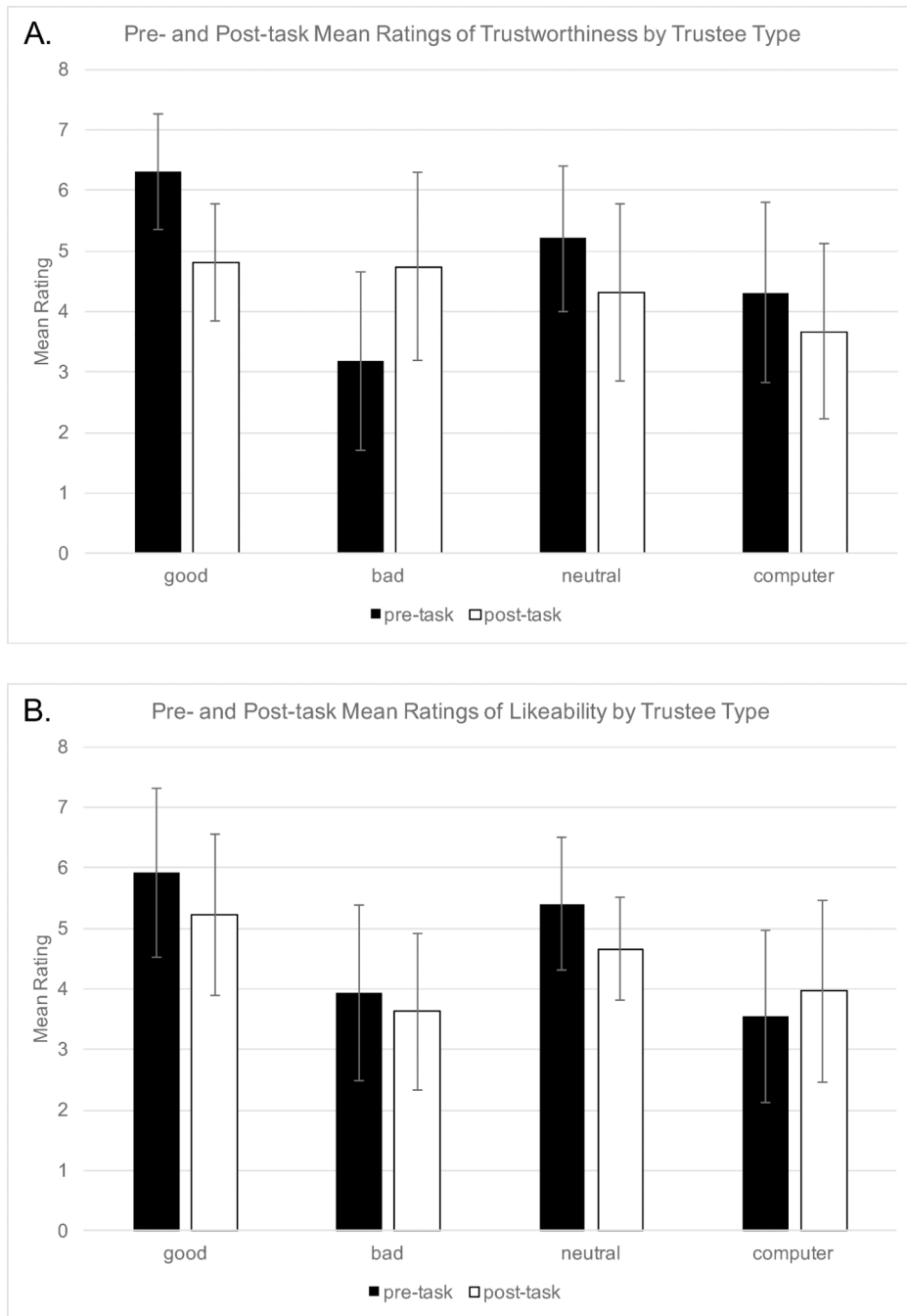
difference between actual rewards and would-be outcome from subject's alternative action. * indicates that results withstand correction for the Bayesian omnibus error rate.

**A.** Pre- and Post-task Mean Ratings of Trustworthiness by Trustee Type

**B.** Pre- and Post-task Mean Ratings of Likeability by Trustee Type

Figure 3.
Manipulation check: pre- and post- task ratings of trustworthiness (A) and likeability (B).
(A) While the good trustee is initially perceived as more trustworthy, and the bad trustee as
less trustworthy, these perceptions revert to the mean following experiential learning. (B) A
similar pattern is seen for likeability, but effects of reputation are more persistent.
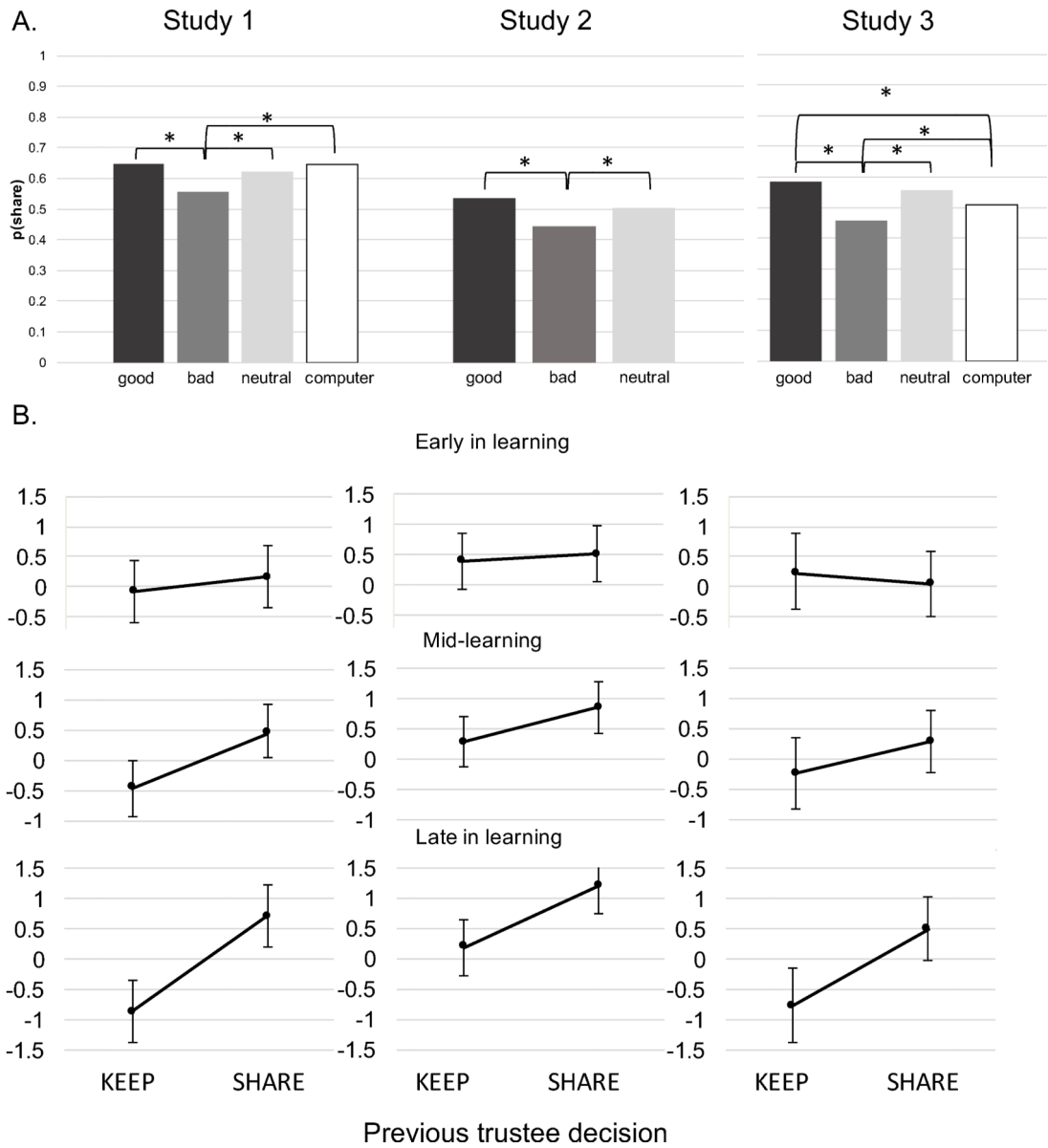
Figure 4.
(A) Observed probability of investing by reputation (trustee type) in Studies 1–3; (B) Predicted likelihood of investing by previous trustee decision and exchange number (previous trustee decision × exchange number), indicating the growing effect of reinforcement as the participant leams (adaptive learning).
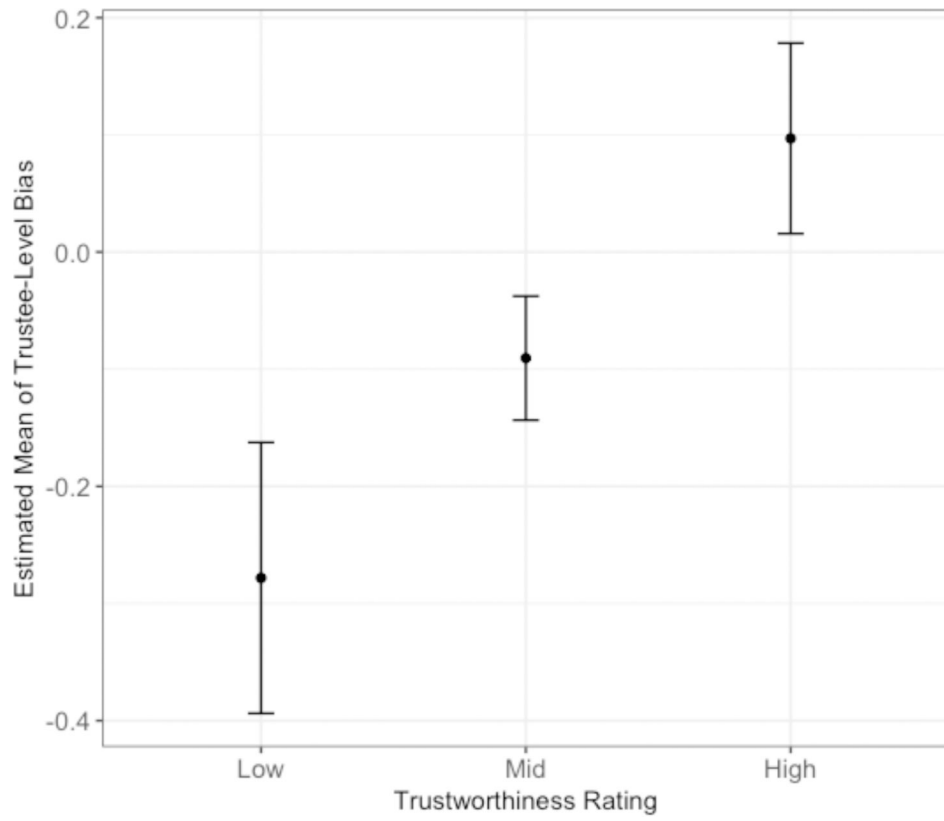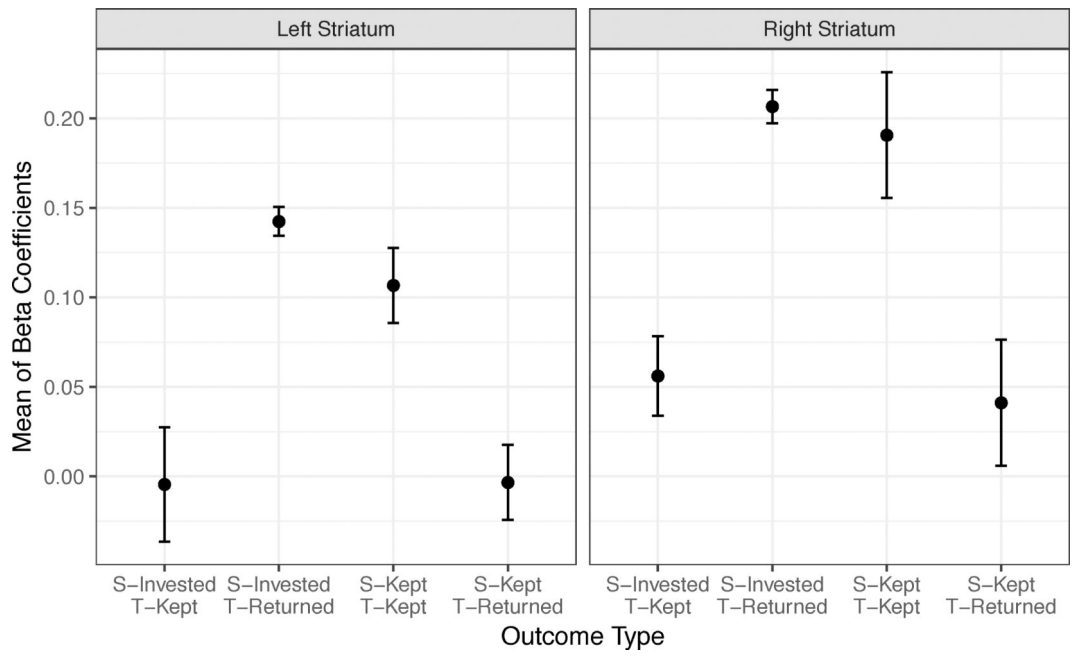
Figure 5.
Estimated means of trustee-level bias parameter by low (1), mid (4), and high (7) values of trustworthiness ratings.
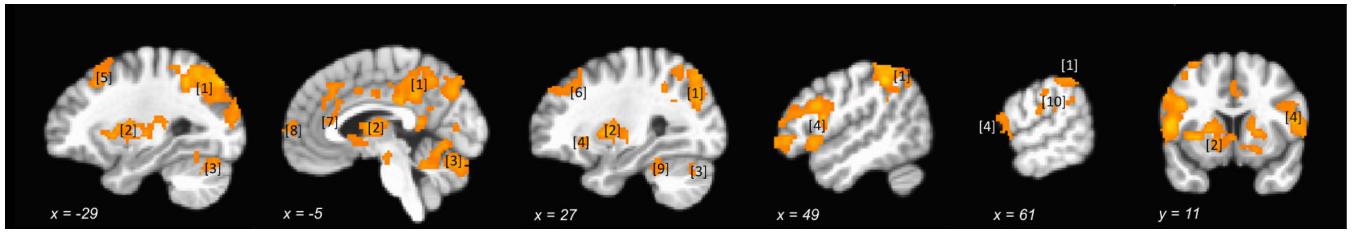
Figure 6.
BOLD response to PEs derived from the policy model ($p_{voxeiwise}$ < .001, cluster size = 100 voxels, p_{corr} < .05).

Figure 7.
Means of β coefficients for the four types of trial outcomes (subject invests/trustee keeps; subject invests/trustee returns; subject keeps/trustee keeps; subject keeps/trustee returns).

Table 1.

Regions representing policy prediction errors.

Region	MNI Coordinates	Peak $t(17)$	Cluster Size (mm^3)
[1] Left/Right Precuneus (BA 7)	−10, −52, 40	8.91	124858
[2] Left/Right Lentiform Nucleus/Putamen	−25, 9, 10	9.47	65422
[3] Left/Right Declive	−10, −69, −22	7.39	21085
[4] Right Inferior Frontal Gyrus (BA 46)	46, 29, 11	7.72	19662
[5] Left Superior Frontal Gyrus (BA 8)	−32, 17, 55	6.06	5706
[6] Right Middle Frontal Gyrus (BA 8)	25, 34, 45	4.87	2275
[7] Left Anterior Cingulate (BA 24/BA 32)	−3, 34, 16	6.21	2227
[8] Left Medial Frontal Gyrus (BA 10)	−6, 65, 8	6.00	1935
[9] Right Culmen	29, −42, −27	5.79	1752
[10] Right Postcentral Gyrus (BA 2)	64, −21, 29	5.64	1679