NeuroImage 174 (2018) 35–43

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Oxytocin attenuates trust as a subset of more general reinforcement learning, with altered reward circuit functional connectivity in males

Jaime S. Ide [a], Sanja Nedic [a], Kin F. Wong [a], Shmuel L. Strey [a], Elizabeth A. Lawson [b,f], Bradford C. Dickerson [c,e,f], Lawrence L. Wald [c,f], Giancarlo La Camera [d], Lilianne R. Mujica-Parodi [a,c,f,*]

[a] *Department of Biomedical Engineering, Stony Brook University School of Medicine, Stony Brook, NY, 11794, USA*
[b] *Neuroendocrine Unit, Massachusetts General Hospital, Boston, MA, 02114, USA*
[c] *Department of Radiology, A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, 02129, USA*
[d] *Department of Neurobiology and Behavior, Stony Brook University, Stony Brook, NY, 11794, USA*
[e] *Department of Neurology, Massachusetts General Hospital, Boston, MA, 02114, USA*
[f] *Harvard Medical School, Boston, MA, 02115, USA*

ARTICLE INFO

Keywords:
Oxytocin
fMRI
Bayesian modeling
Reinforcement learning
Human
Male

ABSTRACT

Oxytocin (OT) is an endogenous neuropeptide that, while originally thought to promote trust, has more recently been found to be context-dependent. Here we extend experimental paradigms previously restricted to *de novo* decision-to-trust, to a more realistic environment in which social relationships evolve in response to iterative feedback over twenty interactions. In a randomized, double blind, placebo-controlled within-subject/crossover experiment of human adult males, we investigated the effects of a single dose of intranasal OT (40 IU) on Bayesian expectation updating and reinforcement learning within a social context, with associated brain circuit dynamics. Subjects participated in a neuroeconomic task (*Iterative Trust Game*) designed to probe iterative social learning while their brains were scanned using ultra-high field (7T) fMRI. We modeled each subject's behavior using Bayesian updating of belief-states ("willingness to trust") as well as canonical measures of reinforcement learning (*learning rate, inverse temperature*). Behavioral trajectories were then used as regressors within fMRI activation and connectivity analyses to identify corresponding brain network functionality affected by OT. Behaviorally, OT reduced feedback learning, without bias with respect to positive versus negative reward. Neurobiologically, reduced learning under OT was associated with muted communication between three key nodes within the reward circuit: the *orbitofrontal cortex, amygdala,* and *lateral (limbic) habenula*. Our data suggest that OT, rather than inspiring feelings of generosity, instead attenuates the brain's encoding of prediction error and therefore its ability to modulate pre-existing beliefs. This effect may underlie OT's putative role in promoting what has typically been reported as 'unjustified trust' in the face of information that suggests likely betrayal, while also resolving apparent contradictions with regard to OT's context-dependent behavioral effects.

Introduction

Oxytocin (OT) is an endogenous neuropeptide that, when exogenously administered intranasally, has been reported to increase people's willingness to trust other humans (Kosfeld et al., 2005), even after betrayal (Baumgartner et al., 2008). The dominant hypothesis is that OT increases trust by reducing fear and associated brain activations in the *amygdala*, *midbrain*, and *dorsal striatum* (Baumgartner et al., 2008). Supporting this hypothesis are findings that OT attenuates the response of the amygdala, as well as that of its related circuits, to fear (Kirsch et al., 2005), conditioned fear (Petrovic et al., 2008), and fearful faces (Domes et al., 2007; Gamer et al., 2010).

The first two studies reporting the behavioral (Kosfeld et al., 2005) and neural (Baumgartner et al., 2008) effects of oxytocin in humans used versions of a neuroeconomic task known as the *Trust Game,* in which player *A* makes a decision about how to split money with player *B*, and then *B* does the same with *A*. Thus, *A*'s split reflects assumptions ('trust') about *B*'s predicted reciprocal behavior. OT increases generosity in the

* Corresponding author. Department of Biomedical Engineering, Stony Brook University, Stony Brook, NY, 11794-5281, USA.
E-mail address: lilianne.strey@stonybrook.edu (L.R. Mujica-Parodi).

https://doi.org/10.1016/j.neuroimage.2018.02.035
Received 11 May 2017; Received in revised form 24 January 2018; Accepted 17 February 2018
Available online 25 February 2018

Trust Game but not in the simpler Dictator Game, the latter of which eliminates assumptions of reciprocity (Zak et al., 2007). Specifically, subjects who were given OT did not change their trusting behavior after receiving information that many trustees had betrayed their trust in previous interactions, whereas subjects who received placebo reduced their trusting behavior after being so informed (Baumgartner et al., 2008). Importantly, in the initial (Kosfeld et al., 2005) study, the effect was reported to be highly trust-specific: oxytocin did not change the behavior of trustees in the Trust Game, nor the behavior of investors in a risky decision task not involving trust. While neuroimaging reports on OT have focused almost exclusively upon the neuropeptide's effect on limbic regions typically associated with fear, one of the earliest fMRI papers on OT showed that it also reduces activity in the *bilateral caudate* (Baumgartner et al., 2008), a key brain structure in reward learning. This raises the question of whether OT's putative effect in blocking the effects of aversive or aversively conditioned stimuli might actually be consequent to a more general diminished recruitment of the reward learning circuit, with associated diminished behavioral adaptation to information feedback.

Previous studies have extensively examined effects of OT in terms of the initial instinct to trust or fear in the absence of (known) prior information (Kirsch et al., 2005; Kosfeld et al., 2005; Baumgartner et al., 2008; Petrovic et al., 2008). However, social relationships typically evolve over time, in response to iterative feedback over the course of many interactions. Therefore, in order to probe the brain circuit dynamics underlying an individual's interaction-evolution, we had subjects play a multi-round version (King-Casas et al., 2005) of the Trust Game (Camerer, 2003) while undergoing 7T fMRI optimized for time-series dynamics at the single-subject level (DeDora et al., 2016). We then modeled their behavior using two approaches.

First, Bayesian modeling described dynamically evolving expectations with regard to positive outcomes. These expectation dynamics were then used as regressors for brain data, to identify neural regions of interest (Yu and Cohen, 2009; Ide et al., 2013) associated with ‘trust’. A subset of these neural regions comprised a reduced functional circuit: *amygdala, nucleus accumbens, orbitofrontal cortex*, previously established by the animal (Dayan and Balleine, 2002), human (O'Doherty et al., 2003), and computational neuroscience (Dayan and Abbott, 2005) literature to underlie reinforcement learning.

Second, we assessed the degree to which subjects (Investors) learned in response to their presumed partners' (Trustees') behavior. This was done using both a simple intuitive measure of previous-trial reciprocity (‘tit-for-tat’), as well as a more rigorous reinforcement learning model quantifying *exploration* (often described as ‘inverse temperature,’ a measure of risk-taking) and *exploitation* (the tendency to capitalize on detected patterns/rules) (Dayan and Abbott, 2005). Using psychophysiological interaction analyses (Gitelman et al., 2003) we then identified condition-specific brain connectivity within the reinforcement learning circuit.

Methods and materials

Subjects and screening procedures

Seventeen healthy male subjects ($\mu_{age} = 25.4 \pm 3.7$ years, $\mu_{weight} = 74 \pm 10$ kg, 2 left-handed) participated in a randomized double-blind within-subject/crossover experiment using a single dose intranasal oxytocin (40 IU) compared to placebo. After an initial phone screening, a study physician obtained written consent from each subject, who then underwent a History and Physical exam. Exclusion criteria included neurological/psychiatric diagnoses, body mass index >30, blood pressure >140/90 mm Hg (or controlled with medication), smoking, and nasal obstruction. Subjects were instructed to abstain from caffeine and alcohol on the day of the scan. Protocols described here were approved by the Institutional Review Boards of Stony Brook University and Partners HealthCare; all subjects provided informed consent.

Administration of oxytocin and placebo

Syntocinon (Oxytocin) Nasal Spray® (Novartis) was administered under FDA IND # 112931. Subjects received 10 sprays (40IU, 1 mL) 60 min prior to the fMRI. Placebo, identical in preparation except for the oxytocin component, was administered in the same manner in a double blind, single-dose, randomized procedure counterbalanced for order. To avoid bleed-through between conditions while controlling for order effects, each session was either oxytocin (OT) only or placebo (PL) only, conducted on separate days; OT and PL were administered at the same time on both days to control for possible diurnal variations in endogenous OT. The number of days between the two sessions ranged between 1 (for 4 out of 17 subjects) and 71, with the median being 7 ($\mu = 14$, s.d. = 20.8).

Studies looking at the effects of intranasal administration of Oxytocin have primarily used a single dose between 24 and 40 IU (Kendrick et al., 2016), with reported dosages ranging from 2 to 40 IU (Wigton et al., 2015), and dose-dependent effects observed in several studies (Cardoso et al., 2013; Quintana et al., 2017), even for lower dosages (Quintana et al., 2015, 2016). The only study to establish that intranasally-delivered neuropeptides do, in fact, cross the blood-brain barrier (Born et al., 2002), used larger doses of a closely related neuropeptide, vasopressin, at 40 and 80IU. They found that CSF concentrations began to rise within 10 min of intranasal administration and continued to increase for up to 80 min after administration. Based upon these results, we chose both the dosage and timing of the study design.

Magnetic resonance imaging

All MRI data were acquired on a 7T Siemens Magnetom scanner (32-channel head-coil array) at the Martinos Center for Biomedical Imaging at MGH. We obtained whole brain EPI BOLD data using parameters previously optimized on this scanner for dynamic fidelity of single-subject time-series (DeDora et al., 2016): SMS slice acceleration factor = 5, GRAPPA acceleration = 2, TR = 802 ms, TE = 20 ms, flip angle = 33°, $2 \times 2 \times 1.5$ mm voxels, 748 measurements (~10 min). Field map images were acquired using: TR = 723 ms, TE1/TE2 = 4.60/5.62 ms, flip angle = 36°, and $1.7 \times 1.7 \times 1.5$ mm voxels. T1-weighted structural volumes were acquired using a conventional MEMPRAGE sequence with 1 mm isotropic voxels and four echoes with TE1/TE2/TE3/TE4 = 1.61/3.47/5.33/7.19 ms, TR = 2530 ms, flip angle = 7°, GRAPPA acceleration = 2.

Iterative trust game

We adapted an iterative version of the Trust Game (King-Casas et al., 2005). During each scan (OT and PL), the subject (“Investor”) played 20 rounds with the same opponent (“Trustee”). At the start of each round, the subject was given 20 monetary units (MU) and told to invest any amount between 0 and 20 with the Trustee. This invested amount was then tripled. The Trustee then repaid some portion of the total (0–60 MU) back to the Investor. While, in reality, the ‘Trustee’ was a computer-generated algorithm, subjects were told they were playing with a human; the deception was revealed following completion of the study.[1] To ensure that the ‘Trustee’ algorithm mimicked human behavior, parameters were estimated from data (N = 48) obtained from a previous study (King-Casas et al., 2005); randomness was set at 10%. As illustrated in Fig. 1a–b, each round consisted of a *cue to invest* (I1), *investment period* (I2), *delay*, *investment reveal* (I3), *delay*, *cue to repay* (R1), *repayment period* (R2), *delay*, *repayment reveal* (R3), *delay*, *totals reveal*, and *inter-round delay*. Delay periods were jittered between 2 and 7s using

[1] During our debriefing, prior to revealing the deception, we asked subjects about their perception of the game. None of the subjects showed evidence of questioning the cover story.

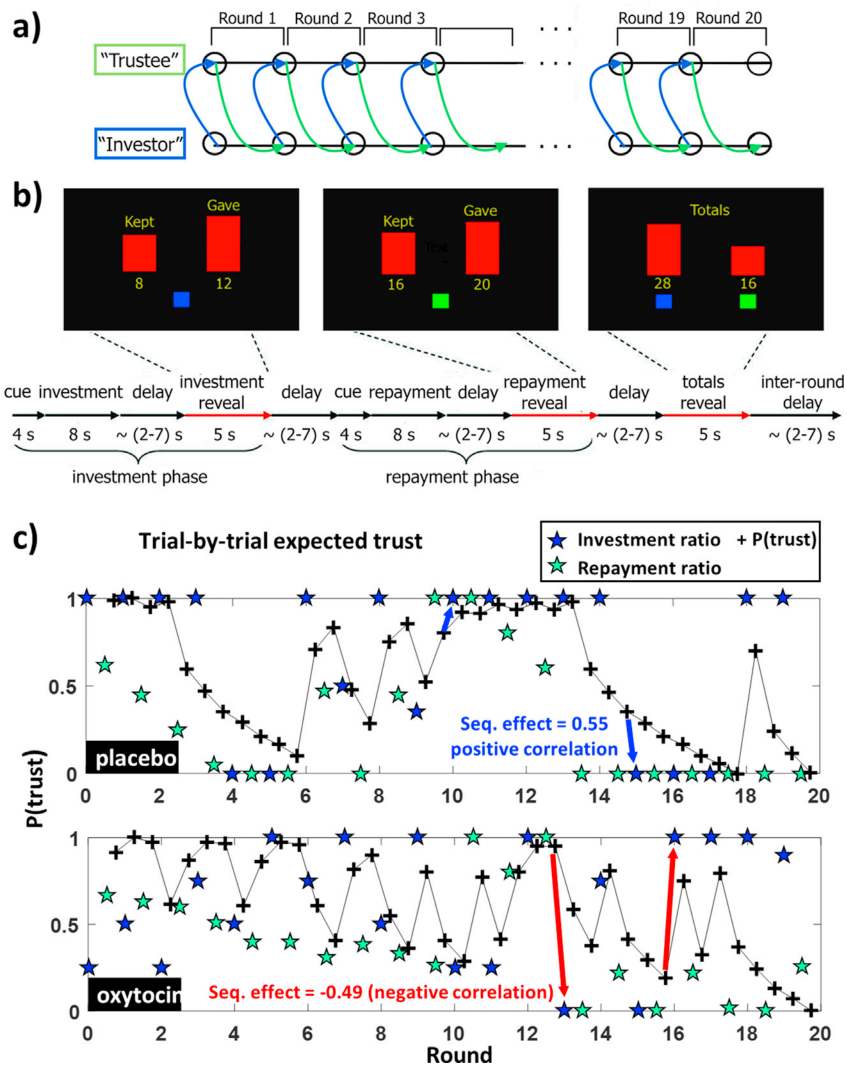

Fig. 1. Repeated trust game. ***(a)*** In this iterative version of the Trust Game, adapted from King-Casas et al. (2005), the subject ('Investor') plays 20 rounds with the same (fictional) opponent ('Trustee'), while being scanned in an MRI scanner. ***(b)*** At the start of each round, the subject is provided 20 monetary units (MU) and told to invest any amount between 0 and 20 with the Trustee. The invested amount is tripled on the way to the Trustee and the Trustee then repays some portion of the total he has (between 0 and 60 MU) back to the Investor. While, subjects were told that they were playing against another person, in fact the 'Trustee' was a computer-generated algorithm (deception was revealed to subjects following completion of the study). ***(c)*** Estimated individual expected values of trust, P(trust), were highly correlated with investment ratio ($r = 0.64$, $p = 0.0032$), as shown for a representative subject for PL but not OT conditions. Aligned or opposite investment ratio and P(trust) across trials are depicted through the blue and red arcs, respectively.

Optseq (https://surfer.nmr.mgh.harvard.edu/optseq/). To increase engagement, MUs partially determined subject compensation; for each session, subjects received a $20 base-payment simply for participating, but then could "earn" up to an additional $30 per session depending upon the number of MUs acquired (payment was: ($20 + MU/10) * 2; thus, total payment for the two sessions ranged from $40-$100). Prior to scanning, subjects were trained by completing two rounds with a designated researcher acting as Trustee.

Evolving belief states: Bayesian expectation of trust

We used a dynamic Bayesian model (Yu and Cohen, 2009; Ide et al., 2013, 2015) to estimate the subject's evolving posterior belief of trust, *P(trust)*. P(trust) was computed for each trial from a *trust signal* that incorporated previous trial history and current observation. We assumed that trust signal $s = 1$ whenever the *investment ratio* (investment divided by 20) was increased or the repayment was larger than investment, with trust signal $s = 0$ otherwise. Subjects were assumed to believe that trial k has probability r_k of signaling trust ($s = 1$), and 1- r_k of not signaling trust ($s = 0$). The Bayesian model assumed that the r_k on trial k has probability θ of being the same as $r_{k\text{-}1}$, and (1-θ) of being re-sampled from a fixed distribution $\pi(r_k)$. Subjects were assumed to use Bayesian inference to update their prior belief of trusting on trial k, $p(r_k|\mathbf{s}_{k\text{-}1})$, based on the prior in the last trial $p(r_{k\text{-}1}|\mathbf{s}_{k\text{-}1})$ and the last trial's true category ($s_k = 1$ for trust signal, $s_k = 0$ otherwise), where $\mathbf{s}_k = [s_1, \ldots, s_k]$ denotes all trials 1 through k. Given the posterior distribution $p(r_{k\text{-}1}|\mathbf{s}_{k\text{-}1})$ on trial k-1, the prior distribution of trust in trial k is given by: $p(r_k|\mathbf{s}_{k\text{-}1}) = \theta\ p(r_{k\text{-}1}|\mathbf{s}_{k\text{-}1}) + (1\text{-}\theta)\ \pi(r_k)$, where the fixed distribution $\pi(r_k)$ is assumed to be a beta distribution with prior mean *pm* and shape parameter "scale" *sc*. The posterior distribution is computed from the prior distribution and the outcome according to the Bayes' rule: $p(r_k|\mathbf{s}_k) \propto p(s_k|r_k)\ p(r_k|\mathbf{s}_{k\text{-}1})$. We defined the Bayesian estimate of trust on trial k, P(trust), as the mean of the predictive distribution $p(r_k|\mathbf{s}_{k\text{-}1})$. We then entered P(trust) as a parametric modulator in general linear model (GLM) analyses (Daw et al., 2006; O'Doherty et al., 2006; Ide et al., 2013) to obtain brain responses linked to dynamic behavioral measures of iterative learning.

The Bayesian model fit was performed for each subject following the optimization procedures presented previously in Ide et al. (2015). In short, we found the optimal set of parameters {theta, prior mean} that produced the highest correlation between the estimated P(trust) and the investment values. The tested values were in the range [0.5 1] and [0 1] for theta and prior mean, respectively. To simplify the search, the scale parameter was set to 10 since it didn't affect the values of P(trust) significantly. There were no significant differences in optimal theta and prior mean values between PL and OT conditions (Wilcoxon rank test, $p > 0.05$). To generate the estimated P(trust) values to be entered as parametric modulator in the GLM, we used the group average optimal parameters, theta = 0.79(±0.21) and pm = 0.37(±0.33).

Learning from feedback: tit-for-tat and reinforcement learning

For the simplest and most intuitive assessment of subjects' learning from feedback, we assessed most-recent-trial reciprocity (*Tit-for-Tat*). We computed the *investment ratio* as the ratio of the actual investment and the maximum allowed amount of 20 units, and analogously for the *repayment ratio*. *Benevolent rounds* were defined as those with increased investment ratios even after a decreased repayment ratio. Conversely, *malevolent rounds* were defined as those with decreased investment ratios even after an increased repayment ratio. These differences (*deltas*) between consecutive ratios were increased or decreased whenever their absolute values were greater than 0.01. Investors' *reciprocity* was defined as the difference between the current investment delta ratio and the previous repayment delta ratio. 'Tit-for-tat' rounds were defined as those with neutral reciprocity (i.e., reciprocity = 0). Investors' *generosity* was defined as the difference between the investment and repayment reciprocities: *generosity*(t) = *investor's reciprocity*(t) – *trustee's reciprocity*(t-1).

For a more rigorous estimate of learning from feedback, we used a standard reinforcement learning model (Dayan and Abbott, 2005), similar to implementations by La Camera and Richmond (2008) and van den Bos et al. (2012). Given a set of possible actions $\{a_1, \ldots, a_n\}$, and a set of associated action values $\{V_1, \ldots, V_n\}$, where n is the number of possible choices for each trial t, we update the action value of the currently selected choice j using the expression: $V_j\,(t) = V_j\,(t\text{-}1) + \alpha(r(t) - V_j\,(t\text{-}1))$. α is the *learning rate*, and the difference $r(t) - V_j\,(t\text{-}1)$ is the *prediction error* between the obtained reward r(t) and the expected action value $V_j(t\text{-}1)$. Greater learning rate α designates greater response to the reward feedback or prediction error. Given the set of action values, we used a *softmax* decision paradigm, for which the probability associated with each choice a_j is computed using a sigmoid function $P(a_j) = \exp(\beta V_j)/\sum_{i=1}^{n}\exp(\beta V_i)$, where β is the inverse temperature (larger βs indicate more deterministic greedy actions). For the Iterative Trust Game, we defined the reward *r(t)* as the total monetary reward in each round *t*; participants learn how much they can trust by building an accurate prediction of the consequences of their actions (low prediction errors). Given a sequence of observed actions and rewards, we found the subject-specific optimal learning rate and inverse temperature that minimized the total negative log-likelihood, computed as the sum of $-\ln P(a_j|model)$ of each observed action a_j (MATLAB *fminsearch*).

Functional MRI preprocessing, activation, and connectivity analyses

Neuroimaging data were preprocessed and analyzed with SPM12 (Wellcome Department of Imaging Neuroscience, University College London, U.K.). Anatomical images were normalized to an MNI template. EPI images were realigned to account for motion, unwarped to correct for distortions caused by magnetic field inhomogeneity, normalized to MNI space, and spatially smoothed (6 mm FWHM). One subject was excluded from all neuroimaging analyses due to severe head motion. Standard general linear models (GLM) were constructed (Friston et al., 1995) using the experimental conditions as main regressors (Investors' I1, I2, I3 and Trustees' R1, R2, R3 conditions), six head-movement parameters as covariates, and *P(trust)*, as well as the reinforcement learning variables (*exploitation* and *prediction error*), as parametric modulators. The contrasts of interest were I2>R2 and R3>I3, for Investors' brains during decision making and after repayment reveal. To delineate functional interactions within the reward and learning networks, we used *psycho-physiological interaction analysis* (PPI) (Gitelman et al., 2003), as implemented in our previous work (Duann et al., 2009; Ide and Li, 2011). To define the seeds, we employed anatomically defined masks for *a priori* regions of interest defined by previous studies of reinforcement learning (Dayan and Abbott, 2005) and oxytocin (Bethlehem et al., 2013): the *amygdala, nucleus accumbens* and the *orbitofrontal cortex* (*OFC*) (Desikan et al., 2006; Zaborszky et al., 2008; Tziortzi et al., 2014) available in FSL (FMRIB Software Library v5.0). These are publicly available masks and thus facilitate replication of the study. Standard volume of interest (VOI) time series extractions were performed by computing the first eigenvariate inside the ROI masks and adjusting for effects of interest (Stephan et al., 2010). These time series constituted the physiological variable and were de-convolved to remove the effects of hemodynamic response function (HRF), multiplied by the psychological variable (contrasts I2>R2 or R3>I3), and re-convolved with the canonical HRF to obtain the interaction term or PPI variable (Gitelman et al., 2003). The three variables were entered as regressors in a whole-brain GLM. PPI analyses were performed for each individual subject, and the resulting positive contrast images (i.e., "1" for the PPI regressors) were used in the random-effect group analyses (Penny et al., 2004).

Results

More than 'trust,' OT attenuates reinforcement learning (with no impact on risk-taking)[2]

Bayesian models, such as P(trust), characterize changing *belief states* (e.g., 'willingness to trust'). In order to assess the impact of OT on individual subjects' evolving expectations throughout the game, we estimated Bayesian iterative behavioral adjustment to sequential feedback (*sequential effects of trust*) by computing data-driven values of P(trust) and correlating them with investment ratios across 20 rounds. Results for a representative subject are illustrated in Fig. 1c. Under placebo, subjects' willingness to trust changed as a function of feedback ($z_{PL} = 0.33 \pm 0.45$, Fisher-z transformed P(trust)). This effect was reduced under OT ($z_{OT} = 0.19 \pm 0.43$) in 12 out of 17 subjects (71%; individual values provided in Table 1a). However, only after removing a subject who showed poor learning (equal to 0.01) for OT condition (Subject #15; see Table 1b), did the difference in sequential effects between OT and PL achieve statistical significance (paired *t*-test, p = 0.03). While the above might appear to weakly confirm OT's popular perception as a "trust drug," in fact, OT did not selectively affect the number of benevolent or malevolent rounds (Table 1c). Thus, *the neuropeptide's effects on behavior were not specifically pro or anti-social.*

We therefore investigated further. Specifically, we asked whether unjustified trust might be the consequence not of pro-social bias in belief, as has been commonly supposed, but rather of a more general failure to encode and/or make use of relevant prior social feedback—not only from negative experiences, but also from positive ones. This we tested using reinforcement modeling (Dayan and Abbott, 2005). Assuming five possible actions (investment values from 0 to 20 equally divided into 5 ranges[3]), we estimated action values and obtained subject-specific learning parameters. Parameter α quantifies the amount of learning from previous interactions: the larger the α, the more rapidly one learns. The inverse temperature, β, quantifies the amount of exploration vs. exploitation. Subjects with larger β show a tendency to stick to actions with larger predicted value, whereas subjects with smaller β show a

[2] All behavioral statistics were obtained using bootstrapping (resampled N = 10,000; repeated 100 times to obtain average).

[3] The use of five bins reflects optimization of fits. We binned the investment values into five ranges (five alternatives) because this model provided the smallest number of learning rates out of range (i.e., >1) in a range from 3 to 6 alternatives. Contrary to standard convention, we originally allowed learning rates >1 to achieve quick value iteration updates. However, this created an oscillatory set of value functions that reflected the observed oscillatory behavioral patterns. These extreme learning rates reflect poor model fit. To address the latter problem, for subjects with learning rates >1 in at least one of the sessions, we improved model fit by generating adaptive quantization. This was obtained by equally binning the investment values into 5 ranges according to the maximum and minimum values. Using this approach, we were able to improve model fit for 3 out of 5 subjects. Learning rates >1 persisted in 2 subjects, and therefore they were excluded from the statistical comparison of learning rate.

Table 1

Single-subject behavior and group statistics for Oxytocin (OT) vs. Placebo (PL), show that OT most strongly affects general reinforcement learning, rather than belief 'willingness to trust', and fail to support either pro or anti-social biases.

(a) The sequential effects of P(trust) were lower under OT ($z = 0.19 \pm 0.43$) as compared to PL ($z = 0.33 \pm 0.45$) conditions in 12 out of 17 subjects (n.s.; however, excluding subject #15, who showed impaired learning under PL, resulted in $p = 0.03$). *(b)* Learning rates (α) were significantly lower under OT ($\alpha = 0.33 \pm 0.27$) as compared to PL ($\alpha = 0.53 \pm 0.3$) conditions ($p = 0.008$). Subjects #8 and #14 presented learning rates greater than 1.0 and were therefore excluded from the statistical comparisons (analyses that included subjects #8 and #14 showed equivalent results: OT ($\alpha = 0.48 \pm 0.37$), PL ($\alpha = 0.77 \pm 0.59$), $p = 0.0007$). *(c)* Average (standard deviation) values across subjects are provided. Importantly, the number of tit-for-tat behaviors was significantly reduced under oxytocin as compared to placebo condition ($p = 0.02$), whereas trust-specific pro-social (asymmetric) deficits, such as the number of benevolent vs. malevolent rounds, were not observed. All statistics reflect bootstrapping (resampled $N = 10{,}000$; repeated 100 times to obtain average).

a. Single-subject ($N = 17$) Bayesian sequential effects of "willingness to trust": P(trust)

S#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
PL	−0.67	0.07	0.26	0.79	−0.14	0.53	0.09	0.64	0.64	0.55	0.86	0.60	0.47	1.00	−0.20	0.34	−0.15
OT	−0.23	−0.09	−0.56	0.18	−0.17	0.86	0.28	0.30	0.34	−0.49	0.51	0.50	0.30	0.84	0.46	0.42	−0.30

b. Single-subject ($N = 17$) reinforcement learning rates

S#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
PL	0.41	0.52	0.35	0.33	0.37	1.00	0.46	n/a	0.85	0.31	0.87	0.29	0.30	n/a	0.01	0.90	0.92
OT	0.03	0.03	0.33	0.35	0.31	0.54	0.03	n/a	0.73	0.01	0.63	0.19	0.03	0.84	0.52	0.78	0.49

c. Other behavioral measures for Iterative Trust Game (group means)

Measure	Placebo	Oxytocin	p-value
Investment ratio	0.54(±0.12)	0.51(±0.14)	0.35
# benevolent rounds	6.76(±2.01)	6.82(±2.30)	0.84
# malevolent rounds	2.82(±1.47)	3.00(±2.06)	0.63
# tit-for-tat rounds	8.29(±2.44)	6.88(±2.23)	0.02*
Reciprocity	−0.04(±0.06)	−0.07(±0.10)	0.07
Generosity	−0.05(±0.09)	−0.04(±0.11)	0.83

tendency to explore potential alternatives to established patterns (risk).

Remarkably, *learning rates were reduced in 13 out of 15 subjects under OT* (87%; paired *t*-test, $p = 0.007$); individual values provided in Table 1b. Analyses excluded subjects #8 and #14 due to learning rates >1.0; however, analyses with these subjects showed equivalent reduction in learning (88%; paired *t*-test, $p = 0.0007$). There were no differences in the inverse temperature between OT and PL conditions (paired *t*-test, $p = 0.17$), showing that OT did not affect the tendency to take action given the current expectation of the next reward.

To confirm that learning rates and the sequential effects of trust were affected by the hypothesized treatment (PL or OT, within-subject factor), and not the session order (between-subject factor), we additionally conducted a mixed design ANOVA. For the reinforcement learning rate, neither session order nor interaction effects showed significant effects ($F = 0.262$, $p = 0.617$, $F = 1.523$, $p = 0.239$, respectively), while the treatment effect did ($F = 9.057$, $p = 0.01$), as per the paired *t*-test results. For the sequential effects of P(trust), none of the session, treatment or interaction effects was significant ($F = 0.710$, $p = 0.415$; $F = 1.223$, $p = 0.289$; $F = 0.217$, $p = 0.649$, respectively).

OT reduces activation of neurobiological network associated with emotional salience, including reduced connectivity within learning circuit[4]

OT Reduces Activation of Salience Circuit During Decision-Making. With respect to the investment period (the period during which the subject's belief-state guides decision-making, as compared to the repayment period), we found that the behavior-driven Bayesian measure of belief updating, P(trust), was significantly coupled to several brain regions (contrast I2>R2 positively modulated by P(trust), one sample *t*-test in the whole group, $p < 0.05$ corrected). These were the *supplementary motor area/middle cingulate cortex* (MCC, $Z = 4.38$, 24 voxels, peak [0 2 52]), the *dorsal anterior cingulate* (ACC, $Z = 4.59$, 35 voxels, peak [2 34 8]), the *left head of caudate* ($Z = 4.26$, 27 voxels, peak [−16 24 0]), and the *left orbitofrontal cortex* (OFC, $Z = 4.14$, 103 voxels, peak [-38 36–14]) (Fig. 2a). Complementary views of these clusters are depicted in Fig. S3 (Supplementary Material). Importantly, increase of P(trust) under OT was associated with a decrease in activation of the *bilateral amygdala* (negative modulation) (Fig. 2b–c), the primary excitatory component of the prefrontal-limbic circuit associated with fear and emotional salience (contrast I1+I2+I3>0 positively modulated by P(trust), paired *t*-test PL > OT, $p < 0.05$ corrected). These results are consistent with previous results from this same task, which found the head of caudate to play a critical role in social decision-making (King-Casas et al., 2005), and link that decision-making to previously-reported (Baumgartner et al., 2008) suppression of the amygdala—with corresponding increases in 'willingness to trust'—following administration of OT.

OT Reduces Amygdala Response to Prediction Error (PE). We then further investigated the amygdala's role with respect to the learning effects noted above. Learning (i.e., updating of beliefs) occurs in response to *prediction error*, an individual's perceived mismatch between expected versus actual outcomes. This potential mismatch appears during the repayment period, when the subject is provided information on actual outcomes, as compared to the investment period, which occurs prior to that information. During PL, prediction error triggered a *bilateral amygdala* response ($Z = 3.48$, 45 voxels, peak [-24 -2 -20]; $Z = 3.62$, 31 voxels, peak [30 4–18], respectively), *an effect that was suppressed following administration of OT* (Fig. 3; contrast R3>I3 modulated by prediction error, paired *t*-test: $p_{\mathrm{Lamy}} = 0.0002$, $p_{\mathrm{Ramy}} = 0.006$).

OT reduces functional connectivity within the reinforcement learning circuit. To assess network effects, we computed psychophysiological interaction (PPI) maps for the repayment > investment (R3>I3) contrast, using the *amygdala*, *nucleus accumbens (NAcc)*, and *OFC* as seed regions. No clear networks were observed with amygdala and NAcc seeds. In contrast, the OFC cluster was significantly correlated with several brain regions, including the *bilateral amygdala* and *lateral (limbic) habenula* (LHb), under the PL condition (Fig. 4a). As shown in Fig. 4b, the degree of connectivity between OFC and the left amygdala and LHb was significantly reduced under OT as compared to PL condition (paired *t*-test, $p < 0.05$). Connectivity between OFC and left amygdala correlated with the number of malevolent rounds (Fig. 4c). We focused on the R3>I3 contrast because of its association with learning. PPI results with the investment period contrast, I2>R2, are shown in Fig. S4 (Supplementary Material).

[4] All neuroimaging statistics were obtained with $p < 0.05$, corrected (3DClustSim, $p = 0.005$, $\alpha = 0.05$).

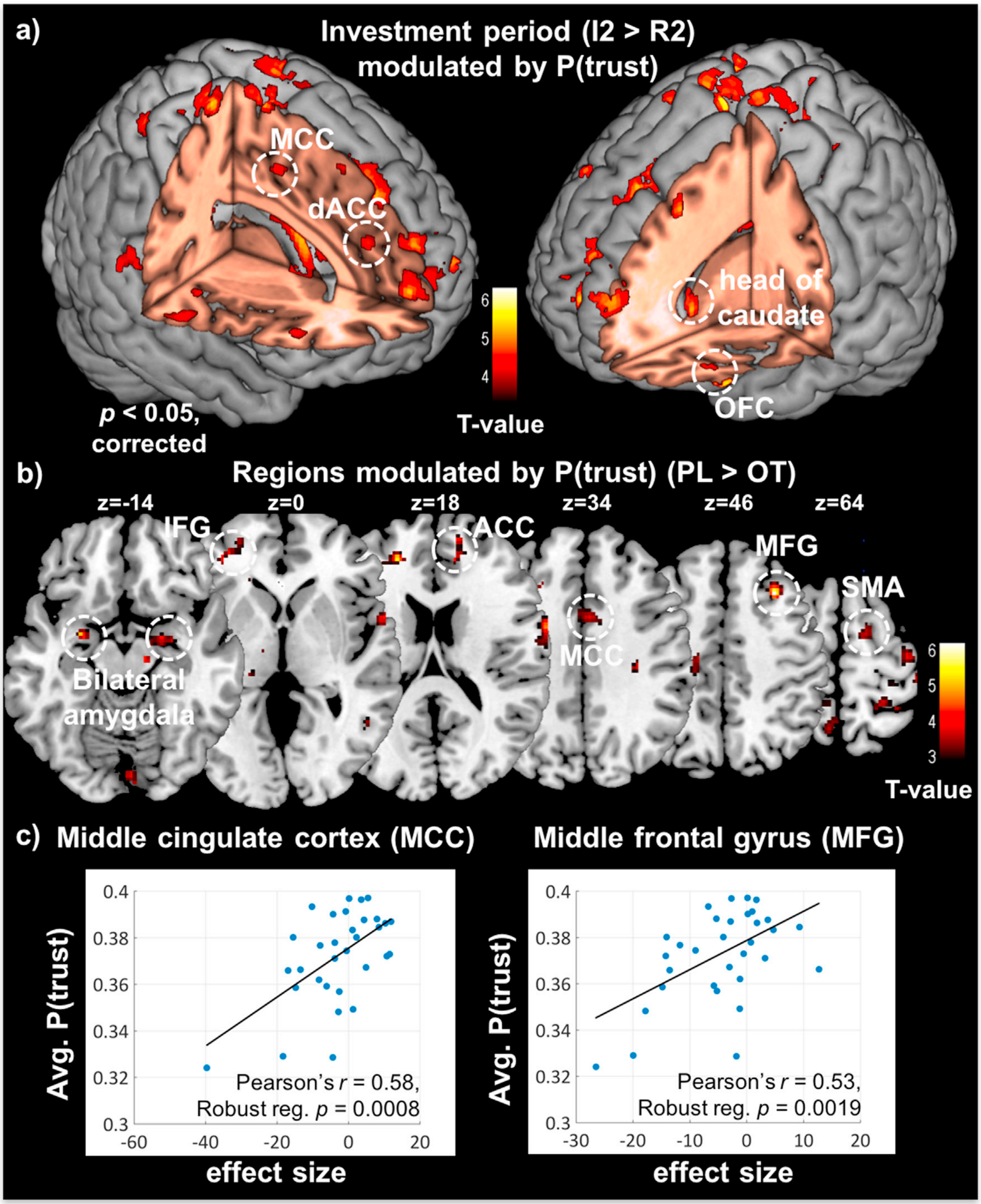

Fig. 2. OT reduces activation of salience circuit during decision-making. ***(a)*** The *head of caudate*, *dorsal anterior cingulate cortex*, *middle cingulate cortex*, and *orbitofrontal cortex* respond to increased P(trust) during investment (one sample *t*-test, $p < 0.05$ corrected, $k > 20$ voxels). ***(b)*** Regions with reduced P(trust) modulation under OT during investment were: the *bilateral amygdala* ($Z = 4.31$, 204 voxels, peak [−26 −2 −14]; $z = 3.9$, 102 voxels, peak [14 −8 −10]), the *inferior frontal gyrus* (IFG, $Z = 3.35$, 83 voxels, peak [−36 42 0]), the *anterior cingulate/medial frontal cortex* ($Z = 3.23$, 43 voxels, peak [12 42 18]), the *middle cingulate cortex* (MCC, $Z = 3.01$, 47 voxels, peak [4 8 34]), and the *middle frontal gyrus* (MFG, $Z = 4.2$, 93 voxels, peak [30 22 46]) (paired *t*-test, $p < 0.05$ corrected). Effect sizes in the bilateral amygdala for PL and OT conditions were 2.55(±1.42) and −6.04(±2.06), respectively. ***(c)*** Increase in P(trust) modulation in the MCC and SFG was correlated with increased P(trust) across subjects.

Discussion

Our data suggest that OT, rather than inspiring feelings of generosity, instead attenuates the brain's encoding of prediction error and therefore its ability to modulate pre-existing beliefs. This effect may underlie OT's putative role in promoting what has typically been reported as 'unjustified trust' in the face of information that suggests likely betrayal. Our research design, which tested modulation of beliefs symmetrically in response to both negative *and* positive experience, demonstrates that OT equally promotes 'unjustified distrust' in the face of information that suggests likely reward. Thus, OT's 'trust' effects would appear to be a subset of a more general attenuation of learning.

In the animal literature, the OFC and amygdala are considered to be canonical regions of the reinforcement learning circuit, while the LHb

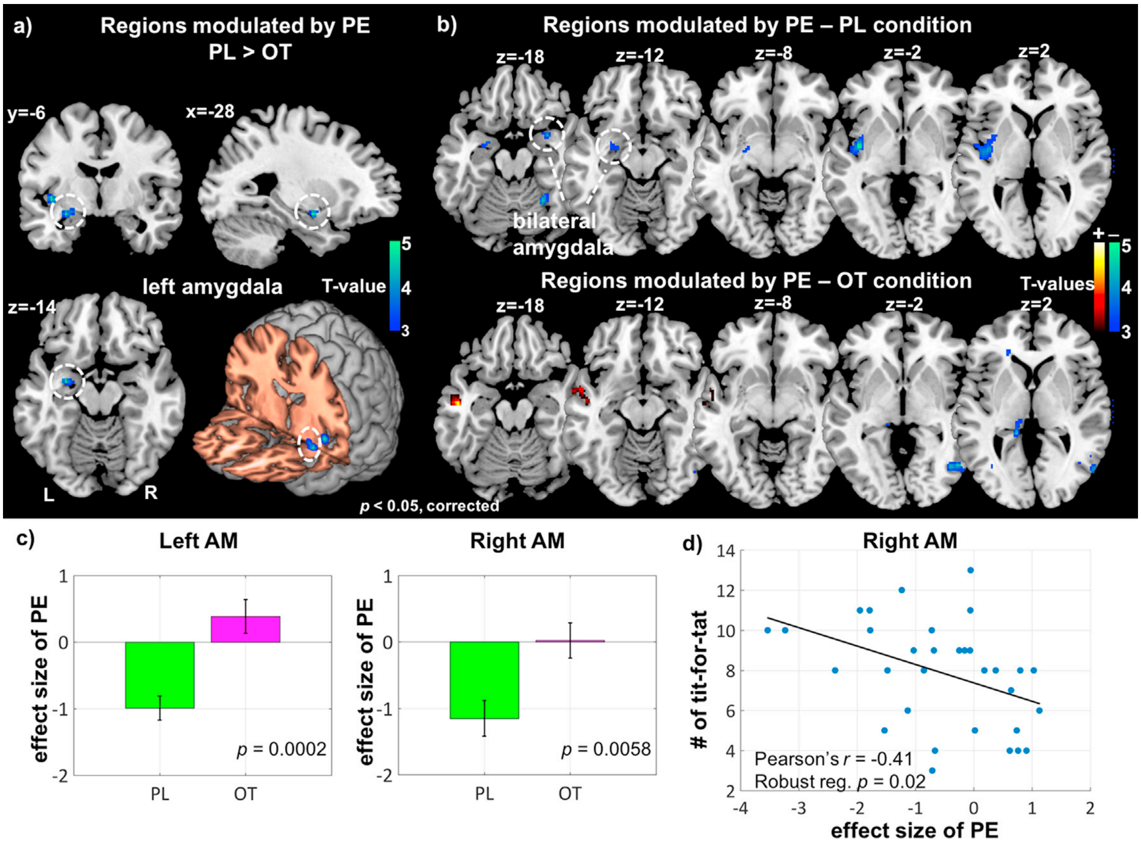

Fig. 3. OT reduces amygdala response to prediction error (PE). ***(a)*** Regions in the brain negatively responding to PE during placebo (PL) as compared to oxytocin (OT) conditions (paired *t*-test, $p < 0.05$ corrected). ***(b)*** Regions responding to PE during placebo PL and OT conditions separately (one-sample t-tests, $p < 0.05$, corrected). ***(c)*** Effect sizes of PE modulation are significantly reduced under PL, but not OT, conditions in the *left* and *right amygdala* (AM) clusters (paired *t*-test, $p = 0.0002$ and $p = 0.006$ respectively). ***(d)*** Effect size of PE modulation in the right AM cluster was negatively correlated with number of tit-for-tat rounds (Pearson's $r = -0.41$, robust regression $p = 0.02$).

acts as a catalyst for the dopaminergic reward response and associated decision-making. The OFC and amygdala mediate social judgment but also basic reward processing (Adolphs, 2003; Viviani et al., 2011; Dölen et al., 2013). Recent work has started to dissect the mechanistic action of OT on these areas. External application of OT, or light stimulation of OT fibers, activates GABAergic neurons in the lateral amygdala that inhibit central amygdala neurons (the main amygdala output) of mice (Viviani et al., 2011). Moreover, OT induces long-term depression (LTD) in the accumbens due to a decrease in presynaptic probability of release from medium spiny neurons, with less LTD observed in mice previously exposed to social conditioning (Dölen et al., 2013). In primates, LHb neurons activate dopamine selectively to punishment and de-activate in response to reward (Matsumoto and Hikosaka, 2007, 2009). These dopaminergic outputs project to the *ventral tegmental area* (Christoph et al., 1986), which in turn projects to the *amygdala*, *cingulate gyrus*, *hippocampus*, *nucleus accumbens*, and *prefrontal cortex* (RC et al., 2009). Behaviorally, rats with this region inactivated by GABA agonists show indifference to 'costs' associated with potential rewards, but not rewards themselves (Stopper and Floresco, 2014). In fact, our result (reduced amygdala response to prediction error during R3>I3, Fig. 3) indicates that OT's effect was not restricted to making use of relevant information during the investment period, but applied also to the encoding of new information required for learning in response to feedback. Therefore, it suggests that OT's behavioral consequences with respect to attenuated learning may be related to compromised communication between different components of the reward circuit associated with detecting and processing prediction error.

One important direction for future research is to investigate whether the fundamentally revised perspective on OT's impact on human neurobiology and cognition suggested above may help to resolve apparent contradictions with regard to OT's behavioral effects. OT is best known as a neuropeptide associated with social and affiliative behavior in humans (Kosfeld et al., 2005; Baumgartner et al., 2008; Delgado, 2008; Averbeck, 2010; Insel, 2010). However, complicating the popular notion of OT as a pro-social 'hormone of love and trust' are recent studies demonstrating that oxytocin actually can also decrease trust, amplifying aggression (Shamay-Tsoory et al., 2009; Bartz et al., 2011a; Grillon et al., 2013; Ne'eman et al., 2016), anxiety to unpredictable threat (Grillon et al., 2013) and anti-social behavior towards unfamiliar individuals (De Dreu et al., 2010). For lactating female mammals, OT is responsible for both selective bonding towards young as well as protective aggression towards outsiders. These effects have also been demonstrated in humans (Hahn-Holbrook et al., 2011), and may be related to OT's role in strengthening feelings of ethnocentrism (De Dreu et al., 2011). It is possible that, far from being a uniquely human cultural construct, tribal notions of social inclusion and exclusion may be grounded in the same biological biases towards genetic similarity found in non-human primates (Morin et al., 1994; Silk, 2002) and which underlie kin selection in evolution theory. If so, our results suggest that OT release could serve as a mechanism for discounting the weight of social learning, thereby permitting innate biases (whether positive or negative) towards new social encounters to dominate.

While the lack of difference in the inverse temperature between OT and PL condition implies that OT did not affect risk-taking, consistent with initial behavioral reports of OT which indicated that risk-taking was unaffected (Kosfeld et al., 2005), this study was not designed to compare social versus non-social learning. Thus, future studies are needed to delineate the effect, to determine if attenuation of learning persists in

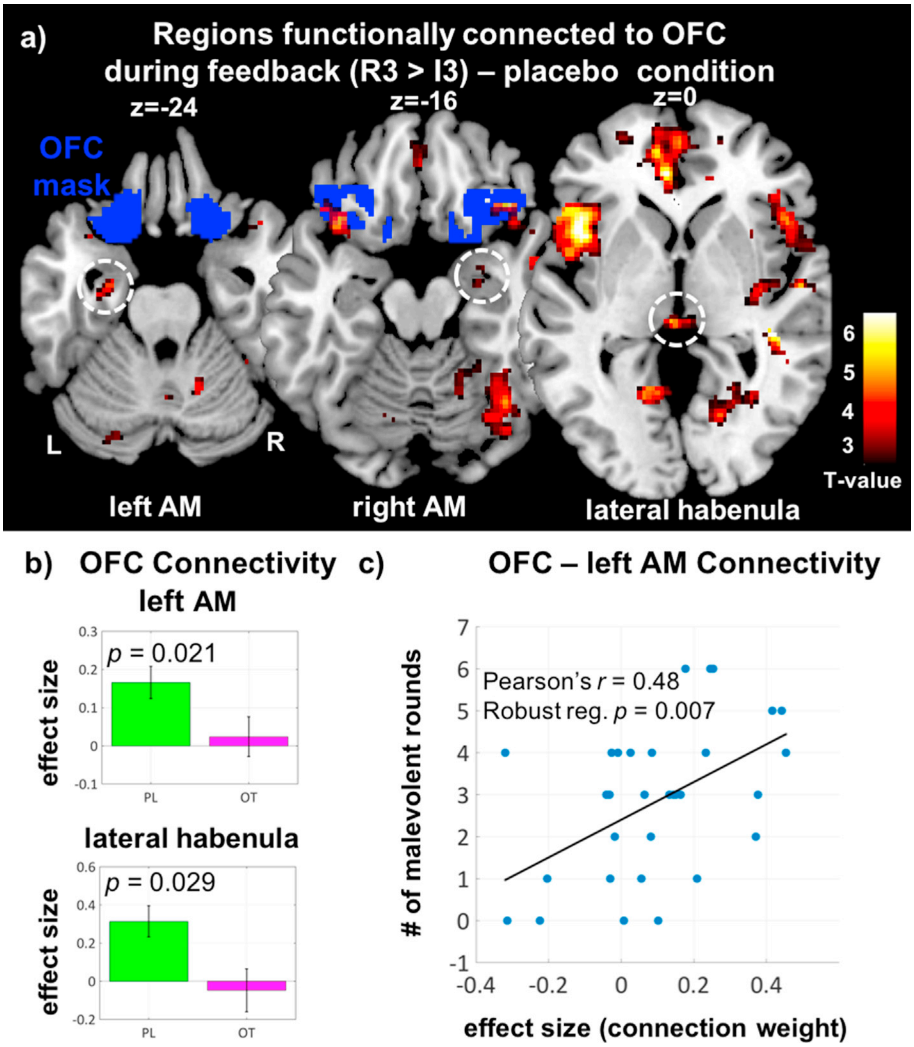

Fig. 4. OT reduces functional connectivity within the reinforcement learning circuit during feedback. ***(a)*** Brain regions connected to OFC during feedback as compared to investment periods (R3>I3) for PL condition. These regions included the *left amygdala* (Z = 3.36, 29 voxels, peak [−26 −12 −26]) and the *bilateral lateral (limbic) habenula* (Z = 3.33, 29 voxels, peak [−4 −26 0]). ***(b)*** Effect sizes of the OFC connectivity with *left amygdala* and *bilateral habenula* were significantly reduced under OT as compared to PL conditions (paired *t*-test, p = 0.021 and p = 0.029, respectively). ***(c)*** Effect sizes for OFC–left AM connectivity were positively correlated with number of malevolent rounds (Pearson's r = 0.48, robust regression p = 0.007).

non-social contexts. Additionally, our findings are limited to men in a specific context in which cooperation is incentivized. Given the likelihood of sex-differences (Rilling et al., 2014), and the fact that a recent study (Lambert et al., 2017), found that OT promoted cooperation in risk-averse women during coordination game, but facilitated aggression in a competitive game, two important future directions will be to see if our effects replicate for women, as well as for adversarial contexts.

Since the seminal discovery that exogenous administration of OT affects human behavior (Kosfeld et al., 2005), nearly 1200 reports have confirmed and diversified its initial role in modulating "trust." Our work is unique within the OT field due to its interpretation of social relationship building through the lens of reinforcement learning, which has a rich history within the fields of computational and basic neuroscience. We chose a task with clear connections to the initial Kosfeld experiment (Trust Study), but used a modification from outside the OT field (King-Casas et al., 2005) to quantify the dynamic process of learning in response to iterative feedback without bias with respect to pro or anti-social effects. It has been proposed (Bartz et al., 2011b) that OT may affect behavior through one of three (possibly complementary) mechanisms: by *reducing anxiety*, by *activating affiliative motivation,* or by *increasing salience of social cues*. Here, our data suggest a fourth option, that OT attenuates social reward learning, a hypothesis that deserves further independent testing with larger sample sizes and both sexes. If this mechanism proves to be correct, however, clinicians will need to carefully consider the implications of the field's early enthusiasm for using OT as a therapeutic intervention for autism, particularly during early-childhood development.

Acknowledgements

This study was funded by the Office of Naval Research #N000141210393 (LRMP and GLC); the authors declare no conflict of interest. The authors thank P. Read Montague for his generous contribution of human behavioral data (N = 48), which were used to develop the algorithms responsible for realistic 'Trustee' behavior in our automated version of the Iterative Trust Study.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.02.035.

References

Adolphs, R., 2003. Cognitive neuroscience of human social behaviour. Nat. Rev. Neurosci. 4, 165–178.
Averbeck, B.B., 2010. Oxytocin and the salience of social cues. Proc. Natl. Acad. Sci. U. S. A. 107, 9033–9034.
Bartz, J., Simeon, D., Hamilton, H., Kim, S., Crystal, S., Braun, A., Vicens, V., Hollander, E., 2011a. Oxytocin can hinder trust and cooperation in borderline personality disorder. Soc. Cognit. Affect Neurosci. 6, 556–563.

Bartz, J.A., Zaki, J., Bolger, N., Ochsner, K.N., 2011b. Social effects of oxytocin in humans: context and person matter. Trends Cognit. Sci. 15, 301–309.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., Fehr, E., 2008. Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. Neuron 58, 639–650.

Bethlehem, R.A., van Honk, J., Auyeung, B., Baron-Cohen, S., 2013. Oxytocin, brain physiology, and functional connectivity: a review of intranasal oxytocin fMRI studies. Psychoneuroendocrinology 38, 962–974.

Born, J., Lange, T., Kern, W., McGregor, G.P., Bickel, U., Fehm, H.L., 2002. Sniffing neuropeptides: a transnasal approach to the human brain. Nat. Neurosci. 5, 514–516.

Camerer, C.F., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton, NJ.

Cardoso, C., Ellenbogen, M.A., Orlando, M.A., Bacon, S.L., Joober, R., 2013. Intranasal oxytocin attenuates the cortisol response to physical stress: a dose-response study. Psychoneuroendocrinology 38, 399–407.

Christoph, G.R., Leonzio, R.J., Wilcox, K.S., 1986. Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. J. Neurosci. Official J. Soc. Neurosci. 6, 613–619.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J., 2006. Cortical substrates for exploratory decisions in humans. Nature 441, 876.

Dayan, P., Balleine, B.W., 2002. Reward, motivation, and reinforcement learning. Neuron 36, 285–298.

Dayan, P., Abbott, L.F., 2005. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. MIT Press.

De Dreu, C.K., Greer, L.L., Van Kleef, G.A., Shalvi, S., Handgraaf, M.J., 2011. Oxytocin promotes human ethnocentrism. Proc. Natl. Acad. Sci. U. S. A. 108, 1262–1266.

De Dreu, C.K., Greer, L.L., Handgraaf, M.J., Shalvi, S., Van Kleef, G.A., Baas, M., Ten Velden, F.S., Van Dijk, E., Feith, S.W., 2010. The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. Science 328, 1408–1411.

DeDora, D.J., Nedic, S., Katti, P., Arnab, S., Wald, L.L., Takahashi, A., Van Dijk, K.R., Strey, H.H., Mujica-Parodi, L.R., 2016. Signal Fluctuation Sensitivity: an improved metric for optimizing detection of resting-state fMRI networks. Front. Neurosci. 10.

Delgado, M.R., 2008. Fool me once, shame on you; fool me twice, shame on oxytocin. Neuron 58, 470–471.

Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980.

Dölen, G., Darvishzadeh, A., Huang, K.W., Malenka, R.C., 2013. Social reward requires coordinated activity of nucleus accumbens oxytocin and serotonin. Nature 501, 179–184.

Domes, G., Heinrichs, M., Glascher, J., Buchel, C., Braus, D.F., Herpertz, S.C., 2007. Oxytocin attenuates amygdala responses to emotional faces regardless of valence. Biol. Psychiatr. 62, 1187–1190.

Duann, J.R., Ide, J.S., Luo, X., Li, C.S., 2009. Functional connectivity delineates distinct roles of the inferior frontal cortex and presupplementary motor area in stop signal inhibition. J. Neurosci. Official J. Soc. Neurosci. 29, 10171–10179.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.

Gamer, M., Zurowski, B., Buchel, C., 2010. Different amygdala subregions mediate valence-related and attentional effects of oxytocin in humans. Proc. Natl. Acad. Sci. U. S. A. 107, 9400–9405.

Gitelman, D.R., Penny, W.D., Ashburner, J., Friston, K.J., 2003. Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. Neuroimage 19, 200–207.

Grillon, C., Krimsky, M., Charney, D.R., Vytal, K., Ernst, M., Cornwell, B., 2013. Oxytocin increases anxiety to unpredictable threat. Mol. Psychiatr. 18, 958–960.

Hahn-Holbrook, J., Holt-Lunstad, J., Holbrook, C., Coyne, S.M., Lawson, E.T., 2011 Oct. Maternal defense: breast feeding increases aggression by reducing stress. Psychol. Sci. 22 (10), 1288–1295. https://doi.org/10.1177/0956797611420729. Epub 2011 Aug 26.

Ide, J.S., Li, C.S.R., 2011. Error-related functional connectivity of the habenula in humans. Front. Hum. Neurosci. 5.

Ide, J.S., Shenoy, P., Yu, A.J., Li, C.S., 2013. Bayesian prediction and evaluation in the anterior cingulate cortex. J. Neurosci. Official J. Soc. Neurosci. 33, 2039–2047.

Ide, J.S., Hu, S., Zhang, S., Yu, A.J., Li, C.S., 2015 Jun 1. Impaired Bayesian learning for cognitive control in cocaine dependence. Drug Alcohol Depend. 151, 220–227. https://doi.org/10.1016/j.drugalcdep.2015.03.021.

Insel, T.R., 2010. The challenge of translation in social neuroscience: a review of oxytocin, vasopressin, and affiliative behavior. Neuron 65, 768–779.

Kendrick, K.M., Guastella, A.J., Becker, B., 2017 Sep 2. Overview of Human Oxytocin Research. Behav. Neurosci. https://doi.org/10.1007/7854_2017_19.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., Montague, P.R., 2005. Getting to know you: reputation and trust in a two-person economic exchange. Science 308, 78–83.

Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., Gruppe, H., Mattay, V.S., Gallhofer, B., Meyer-Lindenberg, A., 2005. Oxytocin modulates neural circuitry for social cognition and fear in humans. J. Neurosci. Official J. Soc. Neurosci. 25, 11489–11493.

Kosfeld, M., Heinrichs, M., Zak, P.J., Fischbacher, U., Fehr, E., 2005. Oxytocin increases trust in humans. Nature 435, 673–676.

La Camera, G., Richmond, B.J., 2008. Modeling the violation of reward maximization and invariance in reinforcement schedules. PLoS Comput. Biol. 4 e1000131.

Lambert, B., Declerck, C.H., Boone, C., Parizel, P.M., 2017. A functional MRI study on how oxytocin affects decision making in social dilemmas: cooperate as long as it pays off, aggress only when you think you can win. Horm. Behav. 94, 145–152.

Matsumoto, M., Hikosaka, O., 2007. Lateral habenula as a source of negative reward signals in dopamine neurons. Nature 447, 1111–1115.

Matsumoto, M., Hikosaka, O., 2009. Representation of negative motivational value in the primate lateral habenula. Nat. Neurosci. 12, 77–84.

Morin, P.A., Moore, J.J., Chakraborty, R., Jin, L., Goodall, J., Woodruff, D.S., 1994. Kin selection, social structure, gene flow, and the evolution of chimpanzees. Science 265, 1193–1201.

Ne'eman, R., Perach-Barzilay, N., Fischer-Shofty, M., Atias, A., Shamay-Tsoory, S.G., 2016. Intranasal administration of oxytocin increases human aggressive behavior. Horm. Behav. 80, 125–131.

O'Doherty, J.P., Buchanan, T.W., Seymour, B., Dolan, R.J., 2006. Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. Neuron 49, 157–166.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., Dolan, R.J., 2003. Temporal difference models and reward-related learning in the human brain. Neuron 38, 329–337.

Penny, W.D., Holmes, A.P., Friston, K.J., 2004. Random-effects analysis. In: Frackowiak KJF, R.S.J., Frith, C., Dolan, R., Friston, K.J., Price, C.J., Zeki, S., Ashburner, J., Penny, W.D. (Eds.), Human Brain Function, second ed. Academic Press, pp. 843–850.

Petrovic, P., Kalisch, R., Singer, T., Dolan, R.J., 2008. Oxytocin attenuates affective evaluations of conditioned faces and amygdala activity. J. Neurosci. Official J. Soc. Neurosci. 28, 6607–6615.

Quintana, D.S., Westlye, L.T., Alnaes, D., Rustan, O.G., Kaufmann, T., Smerud, K.T., Mahmoud, R.A., Djupesland, P.G., Andreassen, O.A., 2016. Low dose intranasal oxytocin delivered with Breath Powered device dampens amygdala response to emotional stimuli: a peripheral effect-controlled within-subjects randomized dose-response fMRI trial. Psychoneuroendocrinology 69, 180–188.

Quintana, D.S., Westlye, L.T., Rustan, O.G., Tesli, N., Poppy, C.L., Smevik, H., Tesli, M., Roine, M., Mahmoud, R.A., Smerud, K.T., Djupesland, P.G., Andreassen, O.A., 2015. Low-dose oxytocin delivered intranasally with Breath Powered device affects social-cognitive behavior: a randomized four-way crossover trial with nasal cavity dimension assessment. Transl. Psychiatry 5 e602.

Quintana, D.S., Westlye, L.T., Hope, S., Naerland, T., Elvsashagen, T., Dorum, E., Rustan, O., Valstad, M., Rezvaya, L., Lishaugen, H., Stensones, E., Yaqub, S., Smerud, K.T., Mahmoud, R.A., Djupesland, P.G., Andreassen, O.A., 2017. Dose-dependent social-cognitive effects of intranasal oxytocin delivered with novel Breath Powered device in adults with autism spectrum disorder: a randomized placebo-controlled double-blind crossover trial. Transl. Psychiatry 7 e1136.

RC, M., EJ, N., SE, H., 2009. Widely projecting systems: monoamines, acetylcholine, and orexin. In: A, S., RY, B. (Eds.), Molecular Neuropharmacology: a Foundation for Clinical Neuroscience, second ed., vols. 154–147. McGraw-Hill Medical, New York, pp. 147–148.

Rilling, J.K., Demarco, A.C., Hackett, P.D., Chen, X., Gautam, P., Stair, S., Haroon, E., Thompson, R., Ditzen, B., Patel, R., Pagnoni, G., 2014. Sex differences in the neural and behavioral response to intranasal oxytocin and vasopressin during human social interaction. Psychoneuroendocrinology 39, 237–248.

Shamay-Tsoory, S.G., Fischer, M., Dvash, J., Harari, H., Perach-Bloom, N., Levkovitz, Y., 2009. Intranasal administration of oxytocin increases envy and schadenfreude (gloating). Biol. Psychiatr. 66, 864–870.

Silk, J.B., 2002. Kin selection in primate groups. Int. J. Primatol. 23, 849–875.

Stephan, K.E., Penny, W.D., Moran, R.J., den Ouden, H.E.M., Daunizeau, J., Friston, K.J., 2010. Ten simple rules for dynamic causal modeling. Neuroimage 49, 3099–3109.

Stopper, C.M., Floresco, S.B., 2014. What's better for me? Fundamental role for lateral habenula in promoting subjective decision biases. Nat. Neurosci. 17, 33–35.

Tziortzi, A.C., Haber, S.N., Searle, G.E., Tsoumpas, C., Long, C.J., Shotbolt, P., Douaud, G., Jbabdi, S., Behrens, T.E., Rabiner, E.A., Jenkinson, M., Gunn, R.N., 2014. Connectivity-based functional analysis of dopamine release in the striatum using diffusion-weighted MRI and positron emission tomography. Cerebr. Cortex 24, 1165–1177.

van den Bos, W., Cohen, M.X., Kahnt, T., Crone, E.A., 2012. Striatum-medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. Cerebr. Cortex 22, 1247–1255.

Viviani, D., Charlet, A., van den Burg, E., Robinet, C., Hurni, N., Abatis, M., Magara, F., Stoop, R., 2011. Oxytocin selectively gates fear responses through distinct outputs from the central amygdala. Science 333, 104–107.

Wigton, R., Radua, J., Allen, P., Averbeck, B., Meyer-Lindenberg, A., McGuire, P., Shergill, S.S., Fusar-Poli, P., 2015. Neurophysiological effects of acute oxytocin administration: systematic review and meta-analysis of placebo-controlled imaging studies. J. Psychiatry Neurosci. 40, E1–E22.

Yu, A., Cohen, J., 2009. Sequential effects: superstition or rational behavior? In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), NIPS 2008, Advances in Neural Information Processing Systems, twenty-first ed. MIT Press, Vancouver, British Columbia, Canada, pp. 1873–1880.

Zaborszky, L., Hoemke, L., Mohlberg, H., Schleicher, A., Amunts, K., Zilles, K., 2008. Stereotaxic probabilistic maps of the magnocellular cell groups in human basal forebrain. Neuroimage 42, 1127–1141.

Zak, P.J., Stanton, A.A., Ahmadi, S., 2007. Oxytocin increases generosity in humans. PLoS One 2 e1128.