Title: Oxytocin modulates third-party sanctioning of selfish and generous behavior within and between groups

Author: Katie Daughters Antony S.R. Manstead Femke S. Ten Velden Carsten K.W. De Dreu

PII: S0306-4530(16)30867-8
DOI: http://dx.doi.org/doi:10.1016/j.psyneuen.2016.11.039
Reference: PNEC 3478

To appear in:

Received date: 28-10-2016
Accepted date: 29-11-2016

Please cite this article as: {http://dx.doi.org/

Oxytocin Modulates Third-Party Sanctioning of Selfish and Generous Behavior Within and Between Groups

Katie Daughters[a], Antony S. R. Manstead[a], Femke S. Ten Velden[b],

Carsten K. W. De Dreu[c,d, *]

[a] School of Psychology, Cardiff University, Cardiff, United Kingdom, CF10 3AT

[b] Department of Psychology, University of Amsterdam, 1018 WB, Amsterdam, The Netherlands

[c] Center for Experimental Economics and Political Decision Making (CREED), University of Amsterdam, 1018 WB Amsterdam, the Netherlands

[d] Social and Organizational Psychology, Leiden University, 2300 RB Leiden, the Netherlands

[*] Corresponding Author: Carsten K.W. De Dreu

Email: c.k.w.dedreu@fsw.leidenuniv.nl

Telephone: +31 (0)6 15056378

Highlights

- We examined the possibility that third-party punishment and reward of others' trust and reciprocation is modulated by oxytocin
- Punishment (reward) was higher for selfish (generous) investors and trustees when investors and/or trustees were in-group rather than outgroup
- Differential treatment of in-group (versus out-group) investors was especially strong when participants received oxytocin rather than placebo
- We conclude that oxytocin contributes to creating and enforcing in-group norms of cooperation and trust.

Abstract

Human groups function because members trust each other and reciprocate cooperative contributions, and reward others' cooperation and punish their non-cooperation. Here we examined the possibility that such third-party punishment and reward of others' trust and reciprocation is modulated by oxytocin, a neuropeptide generally involved in social bonding and in-group (but not out-group) serving behavior. Healthy males and females (N=100) self-administered a placebo or 24 IU of oxytocin in a randomized, double-blind, between-subjects design. Participants were asked to indicate (incentivized, costly) their level of reward or punishment for in-group (outgroup) investors donating generously or fairly to in-group (outgroup) trustees, who back-transferred generously, fairly or selfishly. Punishment (reward) was higher for selfish (generous) investments and back-transfers when (i) investors were in-group rather than outgroup, and (ii) trustees were in-group rather than outgroup, especially when (iii) participants received oxytocin rather than placebo. It follows, first, that oxytocin

leads individuals to ignore out-groups as long as out-group behavior is not relevant to the in-group and, second, that oxytocin contributes to creating and enforcing in-group norms of cooperation and trust.

Key Words – Oxytocin, parochial altruism, competition, endocrinology, economic games

1. Introduction

Humans are social animals and much of their evolutionary success has been attributed to their capacity to cooperate with others in social groups (Axelrod & Hamilton, 1981). Relative to other species, humans are more likely to cooperate with unfamiliar and genetically unrelated others who go on to form cohesive groups (Hill et al., 2011) with distinct, group-serving norms, traditions, and cultural practices (Fehr & Fischbacher, 2004a; Mesoudi, 2016). Indeed no matter how distinct group norms and traditions may be, one common function underlying many of these aspects is to steer group members away from self-interested behavior and towards group-serving, cooperative behavior (Bowles & Gintis, 2011). Accordingly, norm abiding and in-group benefitting behavior is commonly appreciated and sometimes rewarded, whereas norm violations and selfishness are typically frowned upon and often punished (Balliet & Van Lange, 2013).

Group-living provides fitness functionality to its individual members (Darwin, 1873), and it stands to reason that over evolutionary time humans have become biologically prepared for group-serving behavior, such as costly cooperation and norm compliance (Burnham & Johnson, 2005; Fehr & Fischbacher, 2004a). Resonating with this possibility is work linking group-serving behavior to oxytocin, an evolutionarily ancient neuropeptide that is produced in the hypothalamus, and is pivotal in social bond formation and maintenance (Carter, 2014; Donaldson & Young, 2008; Meyer-Lindenberg, Domes, Kirsch, & Heinrichs, 2011).

Synthesized primarily in the paraventricular and supraoptic nuclei of the hypothalamus (Sakamoto et al., 2007), oxytocin acts as a neuromodulator affecting a range of social acts including (i) the cortico-amygdala circuitry to reduce withdrawal from social threat (Domes et al., 2007); and (ii) the "wanting" mesocorticolimbic circuitry promoting (affiliative) approach to positive social targets (Harari-Dahan & Bernstein, 2014; Kemp & Guastella, 2011).

In group-living species such as prairie voles, meerkats, and primates (including humans), elevated oxytocin (following administration or measured from saliva, blood, or urine) is associated with an increased ability to discriminate between familiar and unfamiliar others (Ferguson, Young, & Insel, 2002; Rimmele, Hediger, Heinrichs, & Klaver, 2009), prosocial approach towards those seen as familiar and in-group (as opposed to unfamiliar or outgroup) (De Dreu, Greer, Van Kleef, Shalvi, & Handgraaf, 2011; De Dreu & Kret, 2015; Declerck, Boone, & Kiyonari, 2010), and with enhanced willingness to protect and defend one's group and its territory (Bosch, 2013; De Dreu, Shalvi, Greer, Van Kleef, & Handgraaf, 2012).

Whereas evidence suggests that oxytocin shifts individuals' focus from their self-interests towards those of their group (De Dreu & Kret, 2016), it is unknown whether (and indeed how) oxytocin also modulates the willingness to police and enforce such group-serving behaviors in others, and in particular in one's in-group. In general, such norm-enforcing tendencies are well-documented and functional for group-living, especially within group cooperation. By policing behaviors that are disadvantageous to, or defy the social norms of, the group, group members are kept from straying into selfish or exploitive behavior that endangers the functionality of the group and reduces group efficiency (Gintis, 2000; Gintis, Bowles, Boyd, & Fehr, 2003).

Experimental work on third-party punishment shows that humans engage in such policing and norm-enforcing behavior (Fehr & Fischbacher, 2004b; Nikiforakis & Mitchell, 2014). In these experiments, participants typically witness an exchange between two other

individuals, one of whom is exploiting (or benefitting) the other. Participants are given an endowment that is valuable to them, and allowed to use part or all of this endowment to punish the perpetrator (and sometimes to reward the victim). The participant is not personally involved, and there are no consequences of the observed social exchange, except that extending a punishment (or reward) is personally costly. Accordingly, it is not in the individual's immediate self-interest to punish others for selfishness, or to reward others for their generosity. Nevertheless, there is converging evidence from different lines of research that humans do punish, at personal cost, selfishness in others, and to a lesser extent, reward others for their cooperation and generosity (Fehr & Gächter, 2002; Hu et al., 2016; Nikiforakis & Mitchell, 2014).

Tendencies to police others and third-party punishment can increase within group levels of cooperation and reduce group members' tendencies to defect (Dreber, Rand, Fudenberg, & Nowak, 2008; Egas & Riedl, 2008). Moreover, third-party punishment appears in-group biased: Costly punishments are given more readily when the 'victim' of the selfish behavior is an in-group rather than outgroup member (Baumgartner, Schiller, Rieskamp, Gianotti, & Knoch, 2013; Shinada, Yamagishi, & Ohmura, 2004). Possibly, this reflects that in-groups rather than out-groups have stronger fitness functionality to the individual and costly investments, including those in third-party punishment and reward, have stronger functionality to the individual when targeted at in-group members especially. Here we expected to also find such an in-group bias in third-party decision making. Following the above review on oxytocin, such in-group bias in third-party punishment and reward should be stronger when individuals receive oxytocin (versus placebo).

We examined these possibilities in a novel Third-Party Punishment and Reward Trust Game (TTPR-TG). In this game, participants see the exchange between in investor and a trustee and are can punish or reward the investor and the trustee. Whereas the exchange between

investor and trustee has no consequences to the participant, extending a punishment or reward is costly to the participant and costly (in case of punishment) or rewarding (in case of reward) to the investor or trustee at a 1:3 ratio. The exchange between investor and trustee has the typical properties of a classic Trust Game (Berg et al., 1995): investors have an endowment E from which they can transfer X (with $0 \leq X \leq E$) in their Trustee. Investment X is tripled and Trustees decide on a back-transfer Y (with $0 \leq Y \leq 3X$). In this study, investors and/or trustees were from the participant's in-group or out-group, and manipulated to be generous, fair, or selfish. This allowed us to see whether and how oxytocin influenced punishing and rewarding selfish or generous investors and trustees from one's in-group or out-group.

2. Methods and Materials

2.1 Ethics and Participants

The study was approved by the University of Amsterdam Ethics Committee (file 2015-WOP-4100), and adhered to the Declaration of Helsinki. Participants gave written informed consent prior to the study, and received full debriefing upon completion of the experiment. The study did not involve deception and was fully incentivized.

To estimate the required sample size for this study, we relied on effect sizes reported in earlier studies on oxytocin and in-group bounded cooperation (De Dreu et al., 2010, Experiment 1 and 2, [partial] eta-squared = 0.154 and 0.122, respectively; Ten Velden et al., 2016: [partial] $\eta^2 = 0.048$). Using these observed eta-squared as inputs in G-Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007), with $\alpha = 0.05$ and $\beta = 0.80$, yielded a required sample for between-within interactions in ANOVA of 62, 88, and 108, respectively. Since the last power estimate was based on a study involving male and female subjects, and the current study also targeted a mixed gender sample, we decided to recruit at least 100 participants. This fits the power estimate and sample size of 121 of another recent study in which both male and female subjects participated and following intranasal administration of oxytocin or placebo engaged

in an evaluation/assessment rather than decision-making task (De Dreu, Kret, & Sauter, 2016). In the present study we used stratified sampling to assign healthy males (*N*=36) and females (*N*=64) to conditions, in which they self-administered 24 IU of intranasal OT or an equivalent amount of a matching placebo.

Participants were recruited via an online system, which described the study as investigating the effects of medication on decision making. Participants were offered a monetary reward of €10 for their time, in addition to any earnings accrued during the study. Participants' earnings were determined by decisions made by fellow participants, a fact that was made salient in the instructions. To preserve confidentiality, actual payments were wired by bank transfer to the subject's private account. Per local policy, these transfers could not be made until data collection was complete, as a result, payments were made 3 to 6 weeks after participation.

Exclusion criteria were having a significant physical or psychiatric illness, assessed by medical screening prior to participation. A total of 100 participants took part in the study; one participant was dropped from the analysis due to missing data, leaving 49 participants in the OT condition, and 50 participants in the placebo condition, with a mean age of 21.83 years (SD=3.12). Age did not differ between treatment conditions, $t(97)=.863$, $p > .250$. Due to stratified sampling, the ratio of females to males across treatment was almost identical (OT=32:17 vs. PL=32:18). Female participants' menstrual cycle and oral contraceptive status as self-reported during medical screening (Follicular phase: n = 24; Luteal phase: n = 35; female participants on oral contraceptives: n = 37) did not influence results or conclusions.

2.2 Treatment and Experimental Procedures

The study involved a randomized, double-blind, placebo-controlled, between-subjects design. Participants were asked to refrain from consuming drugs or alcohol the night before the study, and from smoking or drinking caffeine in the 2 hours prior to the study. Using a double-blind

procedure, participants were assigned to either the OT or the placebo condition. They self-administered a placebo or 24 IU (3 puffs of 4 IU per nostril) of Syntocinon (synthetic OT spray, Novartis). The placebo spray matched the OT spray with respect to all ingredients apart from the synthetic OT (De Dreu et al., 2010).

On arrival, participants were seated in individual cubicles so that they could not see or speak to one another. After providing informed consent, they self-administered either a placebo or OT spray under the supervision of the experimenter, who then unlocked the computer; the remainder of the experiment was self-guided (see Figure 1 for a graphical representation of the study's procedure and time-line). In the first 25 minutes, participants completed a series of questionnaires (for further detail see Online Supplementing Information) that had no other function than to fill the a 'wait period' that is typical in OT administration studies both in our own laboratory (Kret & De Dreu, 2013; Shalvi & De Dreu, 2014; Ten Velden, Baas, Shalvi, Kret, & De Dreu, 2014; Ten Velden, Daughters, & De Dreu, 2016), and that of others (Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008; Kirsch et al., 2005; Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). Several studies have demonstrated physiological effects of intranasal administration of OT after this load time (Daughters et al., 2015; Gossen et al., 2012; Van IJzendoorn, Bhandari, Van der Veen, Grewen, & Bakermans-Kranenburg, 2012; Weisman, Zagoory-Sharon, & Feldman, 2012). The computer automatically started the instructions for the experimental tasks and filled out a short questionnaire. This completed the experiment.

Figure 1 About Here

2.3 Materials and Tasks

Participants were randomly allocated to one of two 3-person groups and given instructions for a fully incentivized standard trust game (see Figure 1). It was explained that one fellow-participant (henceforth investor) would be asked to transfer X from their

"Investment Endowment" (IE; with E=€10, and $0 \leq X \leq 10$) to another fellow-participant (henceforth trustee). Transfers would be tripled, and trustees would then be asked to decide a back-transfer amount, Y (with $0 \leq Y \leq 3X$). Next, participants were shown a series of possible exchanges between investors and trustees, and for each exchange they would have the opportunity to assign "evaluation points," first to the investor and then to the trustee (range -10 to +10). Both punishing (negative values, range -10 to -1) and rewarding (positive values, +1 to +10) were costly to the participant: Each point resulted in the subtraction of €0.25 from the participant's "Evaluation Endowment" (EE; with EE=€5) and either subtraction (punishment) or addition (reward) of €0.75 to the target's earnings. Thus, assigning punishing or rewarding points had financial implications for both the participant and the target.

Participants made reward/punishment decisions for both the investor and trustee when the investor was generous, fair, or selfish; and the recipient was generous, fair, or selfish (although some combinations were not presented because they were logically impossible, e.g., investing 0 and back-transferring 20). Because the trustee cannot 'play' in trials where the investor transfers 0, such trials are not included in analyses; as a result, investor transfers are either fair or generous, but trustee back-transfers can be selfish, fair or generous. Thus, we had 11 exchange types: when the investor transferred 0 and the trustee back-transferred 0; when the investor transferred 5 and the trustee back-transferred 0, 5, 10 or 15; and when the investor transferred 10 and the trustee back-transferred 0, 5, 10, 15, 20, or 30 (see Figure 1).

Investors could be in-group (identified by assignment to a team called "Team A", the participant's group) or outgroup (identified by assignment to a team called "Team B", not the participant's group) members, and trustees could also be in-group or outgroup members. Each type of exchange was presented four times (with the investor being in-group or outgroup, and the trustee being in-group or outgroup). Because there were four dyads, two additional members of the group (i.e., A1 and A2), and 11 possible exchange types, for which the

participant had to make two sanctioning decisions (one for the investor and one for the trustee), participants completed 132 sanctioning decisions. For intragroup dyads participants completed 1 trial (and therefore 2 decisions) per exchange type. For intergroup dyads participants completed 2 trials (and therefore 4 decisions) per exchange type and we computed the average across these exchange types.

To incentivize the task and avoid deception, participants completed the experimental task by making six decisions in a standard Trust Game – two as the investor (once playing with an in-group trustee, once playing with an outgroup trustee), and four as the trustee (with a generous [or fair] in-group [or outgroup] investor). As described in the participant's instructions, behavior in the Trust Game would be coupled to a randomly chosen third-party decision trial that matched the exchange in question, so each participant's pay-off was dependent on their own decisions and those by other participants. We explored treatment and gender effects on these trust decisions, and found males to invest more than females (M = 2.57 versus M = 2.28, F(1, 95) = 8.35, p = 0.005) and higher investments in in-group relative to out-group trustees (M = 2.57 versus M = 2.32, F(1, 95) = 17.78, p = 0.0001). No effects for treatment were observed, which corresponds to work showing no overall treatment effects on investment decisions in trust games (Nave, Camerer & McCullough, 2015). For back-transfers, we also found higher back-transfers to in-group rather than out-group members, F(1, 95) = 8.07, p = 0.006, and towards generous rather than fair investors, F(1, 95) = 240.12, p = 0.001.

3. Results

3.1 Data Preparation and Analytic Strategy

Data from one participant were missing due to technical failure and could therefore not be included in the analyses. Exploratory analyses in which we controlled for personality measures (see Supplementary Materials) did not change the results or conclusions.

Hypotheses were tested using a 2 (treatment: oxytocin/placebo) x 2 (gender:

male/female) x 2 (investor group: in-group/outgroup) x 2 (trustee group: in-group/outgroup) x 2 (investor's transfer: generous/fair) x 3 (trustee's back-transfer: generous/fair/selfish) mixed-model Analysis of Variance, with the first two factors being between-subjects, and the remaining factors being within-subjects. Interaction effects were decomposed using simple effects analysis that preserve the overall error term and degrees of freedom (Winer, Brown, & Michels, 1971) (the corresponding test-statistics based on the local error terms and corresponding degrees of freedom replicated results and are reported in brackets where relevant). Accordingly, *p*-values do not have to be corrected for multiple testing as we fitted the data only once, and the most robust estimate of specific contrasts is obtained (Rosenthal & Rosnow, 1985; Tatsuoka & Lohnes, 1988; Winer et al., 1971). Because in several interactions involving within-subjects factors the Mauchly's Test of Sphericity was significant (*p*s < 0.0001) the hypothesis that within-factor error terms are correlated could not be rejected. We thus relied on the more robust yet also more conservative multivariate rather than mixed-model *F*-tests (Tatsuoka & Lohnes, 1988).

Because gender did not interact with treatment in any of the analyses meaning that findings reported below hold across male and female participants. The gender effects we did observe are summarized in the Supplementary Materials. Also, as a robustness check we explored whether findings reported below change when we controlled for personality measures (see Supplementary Materials), but never found this to be the case.

3.2 Third-Party Sanctioning of Investors

Generous transfers were rewarded more (M = 5.587, SE = 0.317) than fair transfers (M = 2.752, SE = 0.251), $F(1, 97) = 100.399$, $p < 0.001$, $\eta_p^2 = 0.514$. Investors received higher rewards when their trustee's back-transfers were selfish (M = 4.827, SE = 0.240) rather than fair (M = 4.131, SE = 0.266) or generous (M = 3.550, SE = 0.287), $F(1.47, 139.42) = 31.724$, $p < 0.001$, $\eta_p^2 = 0.250$, and in-group investors were rewarded more (M = 4.523, SE = 0.256) than outgroup

investors (M = 3.816, SE = 0.264), $F(1, 97) = 15.58$, $p \leq 0.001$, $\eta^2 = 0.138$.

Figure 2 About Here

The main effect for investor group membership was qualified by an interaction with treatment, $F(1, 95) = 6.101$, $p = 0.015$, $\eta^2 = 0.059$ (see Figure 2). Simple effects analysis revealed that whereas participants in the placebo condition did not discriminate between in-group and outgroup investors, $F(1, 95) = 1.10$, $p = 0.297$, those who received oxytocin rewarded in-group investors more than outgroup investors, $F(1, 95) = 20.385$, $p \leq 0.001$. The treatment x investor's group membership interaction was further qualified by an interaction between treatment, investor's group membership, trustee's group membership, and trustee's back-transfer, $F(2, 94) = 4.593$, $p = 0.012$, $\eta^2 = 0.087$. In keeping with the nature of the treatment x investor's group membership effect, we decomposed this complex effect using simple effects analysis for the (interactions among) treatment, investor's group membership, and trustee's group membership within each level of the trustee's back-transfer. Because effects were tested three times (for each level of back-transfer), we corrected for multiple comparisons by setting $\alpha = 0.05/3 = 0.015$ as the critical p-value.

When the trustee's back-transfers were selfish, in-group investors received more when their selfish trustee was outgroup rather than in-group (investor's x trustee's group membership, $F[1, 95] = 6.73$, $p = 0.001$). Furthermore, in-group investors were rewarded more than outgroup investors, $F(1, 95) = 14.71$, $p = 0.001$, $\eta^2 = 0.226$, yet only when subjects received oxytocin, $F(1, 95) = 15.85$, $p = 0.001$ ($F[1,48] = 13.99$, $p = 0.001$), and not when they received placebo, $F(1, 95) = 0.24$, $p = 0.625$ ($F[1,47] = 0.28$, $p = 0.601$) (Fig 3A; overall investor's group membership x treatment, $F[1, 95] = 8.18$, $p = 0.005$).

Figure 3ABC About Here

When the trustee's back-transfers were fair, the interaction among investor and trustee's group membership was not significant, $F < 1$. However, as with selfish back-transfers, in-group

investors were rewarded more than outgroup investors, $F(1, 95) = 11.78$, $p = 0.001$, again only when subjects received oxytocin, $F(1, 95) = 9.32$, $p = 0.003$ ($F[1,48] = 7.53$, $p = 0.009$), and not when they received placebo, $F(1, 95) = 0.79$, $p = 0.376$ ($F[1,47] = 1.03$, $p = 0.315$) (Fig 3B; overall investor's group membership x treatment, $F[1, 95] = 3.84$, $p = 0.053$; marginal).

When the trustee's back-transfers were generous, the interaction among investor and trustee's group membership was again not significant, $F < 1$. Yet here too, in-group investors were rewarded more than outgroup investors, $F(1, 95) = 17.05$, $p = 0.001$, when subjects received oxytocin, $F(1, 95) = 18.18$, $p = 0.001$ ($F[1,48] = 15.53$, $p = 0.001$), and not when they received placebo, $F(1, 95) = 1.84$, $p = 0.178$ ($F[1,47] = 2.21$, $p = 0.143$) (Fig 3C; overall investor's group membership x treatment, $F[1, 95] = 5.40$, $p = 0.022$).

Taken together, subjects who received oxytocin rather than placebo rewarded in-group investors more than outgroup investors. This treatment x investor's group membership interaction was nominally significant at all three levels of the trustee's back-transfer, yet strongest (and surpassed the Bonferonni-corrected threshold) when back-transfers were selfish. Thus, when trustee's back-transfers were selfish, subjects given oxytocin compensated their investors but less so when investors were from the outgroup.

3.3 Third-Party Sanctioning of Trustees

We found that outgroup trustees were rewarded less ($M = 0.989$, SE = 0.207) than in-group trustees ($M = 1.623$, SE = 0.215), $F(1, 95) = 11.51$, $p = 0.001$, $\eta^2 = 0.108$, generous transfers were rewarded more ($M = 2.161$, SE = 0.213) than fair transfers ($M = 0.452$, SE = 0.195), $F(1, 95) = 124.90$, $p < 0.001$, $\eta^2 = 0.568$, and generous back-transfers were rewarded more ($M = 5.914$, SE = 0.304) than fair back-transfers ($M = 2.219$, SE = 0.298), which were rewarded more than selfish back-transfers ($M = -4.213$, SE = 3.29), $F(2, 94) = 199.687$, $p < 0.001$, $\eta^2 = 0.753$. These main effects were qualified in two two-way interactions among investor's group membership and trustee's back-transfer, $F(2, 94) = 8.397$, $p = 0.001$, $\eta^2 = 0.113$, and trustee's

group membership and trustee's back-transfer, $F(2, 94) = 4.34$, $p = 0.016$, $\eta^2 = 0.062$. These were further qualified in two three-way interactions among investor's group membership, trustee's back-transfer, and treatment, $F(2, 94) = 3.074$, $p = 0.051$, $\eta^2 = 0.049$, and trustee's group membership, trustee's back-transfer, and treatment, $F(2, 94) = 2.586$, $p = 0.081$, $\eta^2 = 0.036$ (marginal) and in a four-way interaction among investor's group membership, trustee's group membership, trustee's back-transfer, and treatment, $F(2, 94) = 3.850$, $p = 0.025$, $\eta^2 = 0.062$.

As with results for sanctioning of investors, we probed the nature of these effects with simple effects for (interactions among) investor's group membership, trustee's group membership, and treatment within each level of the trustee's back-transfer. Because simple main and interaction effects were estimated three times (for each level of back-transfer), we corrected for multiple comparisons by setting $\alpha = 0.05/3 = 0.015$ as the critical p-value.

When back-transfers were selfish, selfish trustees were punished more when they were from the outgroup rather than from the in-group, $F(1, 95) = 13.10$, $p = 0.001$, $\eta^2 = 0.115$. Furthermore, trustees were punished more when their investor was from the in-group, rather than the outgroup, $F(1, 95) = 8.63$, $p = 0.004$, $\eta^2 = 0.084$. Although the treatment x investor's group membership was not significant, $F(1, 95) = 1.58$, $p = 0.212$, it can be seen in Fig 4A that effects of investor's group membership were strong and significant when subjects received oxytocin, $F(1, 95) = 9.28$, $p = 0.003$ ($F[1, 48] = 6.87$, $p = 0.012$), rather than placebo, $F(1, 95) = 1.35$, $p = 0.247$ ($F[1, 47] = 2.06$, $p = 0.157$).

Figure 4ABC About Here

When back-transfers were fair, in-group trustees were rewarded more than outgroup trustees, $F(1, 95) = 10.47$, $p = 0.002$, $\eta^2 = 0.086$; there was some evidence that this effect was particularly strong when investors were in-group and subjects received placebo, and when investors were outgroup and subjects received oxytocin, $F(1, 95) = 4.43$, $p = 0.038$, $\eta^2 = 0.027$

(Fig 4B). However, the effect falls above the Bonferroni-corrected threshold and none of the underlying contrasts were (Bonferroni-corrected) significant. We refrain from further interpreting this finding.

When back-transfers were generous, trustees were rewarded more when they faced in-group rather than outgroup investors, $F(1, 95) = 12.22$, $p = 0.001$, but only when subjects received oxytocin, $F(1, 95) = 17.24$, $p = 0.001$ ($F[1,48] = 9.90$, $p = 0.003$), and not when subjects received placebo, $F(1, 95) = 0.17$, $p = 0.680$ ($F[1,47] = 0.63$, $p = 0.434$) (Fig 4C; overall investor group membership x treatment, $F[1, 95] = 6.86$, $p = 0.010$, $\eta^2 = 0.068$).

Taken together, trustees were sanctioned less when they were from the in-group, and when interacting with an in-group rather than outgroup investors. Oxytocin modulated this when back-transfers were generous and, to a lesser extent, when they were selfish. When trustee's back-transfers were generous, subjects given oxytocin rewarded these trustees but less so when the generously treated investors were from the outgroup.

4. Conclusions and Discussion

Individuals in groups punish those who are selfish towards other group members, and reward those who are generous. Such sanctioning may be costly to the individual, yet also functions to sustain cooperative behavior and reciprocity within the group. Indeed, there is some evidence that both punishment of norm violators and rewarding strong contributors is oriented more towards individuals belonging to one's in-group, than towards an outgroup (Baumgartner et al., 2013; Schiller, Baumgartner, & Knoch, 2014; Shinada et al., 2004).

Here we show that such intergroup discrimination in punishment and reward emerges especially when individuals were given oxytocin. Results suggest that individuals given oxytocin (versus placebo) are more likely to (i) condition their punishment and reward decisions on whether the target was in-group or outgroup and (ii) refrain from spending their money on sanctioning outgroup members—oxytocin appears to make people relatively

indifferent about the behavior and fate of outgroup members. Thus, compared to those given placebo, individuals given oxytocin rewarded someone who demonstrated generosity towards another group member, or reciprocated a generous offer, more when the individual was an in-group rather than outgroup member. Likewise, they punished someone who was selfish, or returned a selfish offer, more when the target individual betrayed an in-group rather than an outgroup member.

Findings provide further evidence that oxytocin does not make people ubiquitously more pro-social. This would have manifested in overall reduced punishment and increased reward under oxytocin, which we did not observe. In addition, current findings clarify that oxytocin does not induce unconditional pro-social treatment of in-group members. This would have manifested in reduced punishment and increased rewarding of in-group members only, something we did not observe. Instead, current findings support the idea that the function of third-party sanctioning is to regulate people's adherence to group norms (Chavez & Bicchieri, 2013; Fehr & Fischbacher, 2004b), that oxytocin enhances intergroup discrimination (Ten Velden et al., 2014), and that oxytocin shifts the individual's focus towards the in-group (De Dreu, 2012; De Dreu & Kret, 2015; Ten Velden et al., 2016). Findings also demonstrate that this 'oxytocin shift' is activated in a minimal group paradigm, providing support for the evolutionary function of oxytocin in group behavior via a biological mechanism. It follows that the same pattern of results should also be observed in a real groups paradigm, a potential study for future research.

Our conclusions about the possible role of oxytocin on third-party punishment and reward derive from an exogenous administration study, in which subjects received intranasal oxytocin or a matching placebo. The advantage of this method is that it permits conclusions about cause and effect relationships, which would not be possible if one were correlating endogenous oxytocin with third-party decision-making. The disadvantage of this method,

however, is that the neurophysiological pathways through which intranasal oxytocin affects brain activity and behavioral responses are not fully understood. Although there is good evidence that intranasal oxytocin increases the concentration of endogenous oxytocin found in blood plasma and saliva (Daughters et al., 2015; Gossen et al., 2012; Weisman et al., 2012), the evidence that intranasal oxytocin crosses the blood-brain barrier is limited (Neumann, Maloumby, Beiderbeck, Lukas, & Landgraf, 2013; Paloyelis et al., 2014; Striepens et al., 2013).

In addition to a direct effect on the brain, intranasal oxytocin may also affect brain and behavioral responses through its peripheral effects on the body (e.g., by affecting heart rate, or cortisol responses). Detailing these pathways is an important question for future research. Such new research may also consider the notion that individual differences exist in the peripheral responses to intranasal administration of oxytocin (Daughters et al., 2015). For example, one could investigate whether such individual differences, in turn explain individual differences in the extent to which people pursue in-group bounded cooperation and uphold and enforce in-group serving norms. Future research could also consider adding a greater number of trials for each exchange type and dyad to further improve the reliability of the findings.

Punishing group members for violating norms, acting selfishly, and failing to cooperate with others, serves the same function as rewarding members for upholding norms, being generous, and cooperating with others: It promotes group functioning and effectiveness, and therefore provides indirect benefits for the individual. It follows therefore that humans should be more inclined to punish and reward in-group members, than outgroup members. The current study finds that such functional differentiation between in-group and outgroup members was stronger when individuals received oxytocin. We suggest that oxytocin provides a neurobiological mechanism underlying in-group serving behaviors.

5. References

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(4489), 1390-1396.

Balliet, D., & Van Lange, P. A. (2013). Trust, conflict, and cooperation: a meta-analysis. *Psychological Bulletin, 139*(5), 1090.

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*(4), 639-650.

Baumgartner, T., Schiller, B., Rieskamp, J., Gianotti, L. R., & Knoch, D. (2013). Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction. *Social cognitive and affective neuroscience*, nst023.

Bosch, O. J. (2013). Maternal aggression in rodents: brain oxytocin and vasopressin mediate pup defence. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 368*(1631), 20130085.

Burnham, T. C., & Johnson, D. D. (2005). The biological and evolutionary logic of human cooperation. *Analyse & Kritik, 27*(2), 113-135.

Carter, S. C. (2014). Oxytocin pathways and the evolution of human behavior. *Annual review of psychology, 65*, 17-39.

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology, 39*, 268-277.

Daughters, K., Manstead, A. S., Hubble, K., Rees, A., Thapar, A., & Goozen, H. M. (2015). Salivary oxytocin concentrations in males following intranasal administration of oxytocin: A double-blind, cross-over study. *PLOS one, 10*(12). doi: 10.1371/journal.pone.0145104

Darwin, Ch, (1873). *Origins of species*. Publisher.

De Dreu, C. K. (2012). Oxytocin modulates cooperation within and competition between groups: an integrative review and research agenda. *Hormones and Behavior, 61*(3), 419-428.

De Dreu, C. K., Greer, L. L., Handgraaf, M. J., Shalvi, S., Van Kleef, G. A., Baas, M., . . . Feith, S. W. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science, 328*(5984), 1408-1411. doi: 10.1126/science.1189047

De Dreu, C.K.W., Greer, L. L., Van Kleef, G. A., Shalvi, S., & Handgraaf, M. J. (2011). Oxytocin promotes human ethnocentrism. *Proceedings of the National Academy of Sciences, 108*(4), 1262-1266. doi: 10.1073/pnas.1015316108

De Dreu, C. K.W., & Kret, M. E. (2016). Oxytocin conditions intergroup relations through upregulated in-group empathy, cooperation, conformity, and defense. *Biological Psychiatry, 79*(3), 165-173.

De Dreu, C. K. W., Kret, M. E., & Sauter, D. A. (2016). Assessing emotional vocalizations from cultural in-group and out-group depends on oxytocin. *Social Psychological and Personality Science*, 1948550616657596.

De Dreu, C. K. W., Shalvi, S., Greer, L. L., Van Kleef, G. A., & Handgraaf, M. J. (2012). Oxytocin motivates non-cooperation in intergroup conflict to protect vulnerable in-group members. *PLOS one, 7*(11). doi: 10.1371/journal.pone.0046751

Declerck, C. H., Boone, C., & Kiyonari, T. (2010). Oxytocin and cooperation under conditions of uncertainty: the modulating role of incentives and social information. *Hormones and Behavior, 57*(3), 368-374.

Domes, G., Heinrichs, M., Gläscher, J., Büchel, C., Braus, D. F., & Herpertz, S. C. (2007). Oxytocin attenuates amygdala responses to emotional faces regardless of valence. *Biological Psychiatry, 62*(10), 1187-1190. doi: 10.1016/j.biopsych.2007.03.025

Donaldson, Z. R., & Young, L. J. (2008). Oxytocin, vasopressin, and the neurogenetics of sociality. *Science, 322*(5903), 900-904.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*(7185), 348-351.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of

cooperation. *Proceedings of the Royal Society of London B: Biological Sciences, 275*(1637), 871-878.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175-191.

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*(4), 185-190.

Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63-87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137-140.

Ferguson, J. N., Young, L. J., & Insel, T. R. (2002). The neuroendocrine basis of social recognition. *Frontiers in Neuroendocrinology, 23*(2), 200-224. doi: 10.1006/frne.2002.0229

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of theoretical biology, 206*(2), 169-179.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24*(3), 153-172.

Gossen, A., Hahn, A., Westphal, L., Prinz, S., Schultz, R., Gründer, G., & Spreckelmeyer, K. (2012). Oxytocin plasma concentrations after single intranasal oxytocin administration–A study in healthy men. *Neuropeptides, 46*(5), 211-215.

Harari-Dahan, O., & Bernstein, A. (2014). A general approach-avoidance hypothesis of oxytocin: accounting for social and non-social effects of oxytocin. *Neuroscience & Biobehavioral Reviews, 47*, 506-519.

Hill, K. R., Walker, R. S., Božičević, M., Eder, J., Headland, T., Hewlett, B., . . . Wood, B. (2011). Co-residence patterns in hunter-gatherer societies show unique human social structure. *Science, 331*(6022), 1286-1289.

Hu, Y., Scheele, D., Becker, B., Voos, G., David, B., Hurlemann, R., & Weber, B. (2016). The effect of oxytocin on third-party altruistic decisions in unfair situations: An fMRI study. *Scientific reports, 6*, 20236.

Kemp, A. H., & Guastella, A. J. (2011). The role of oxytocin in human affect: A novel hypothesis. *Current Directions in Psychological Science, 20*(4), 222-231.

Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., . . . Meyer-Lindenberg, A. (2005). Oxytocin modulates neural circuitry for social cognition and fear in humans. *The Journal of Neuroscience, 25*(49), 11489-11493.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673-676.

Kret, M. E., & De Dreu, C. K. (2013). Oxytocin-motivated ally selection is moderated by fetal testosterone exposure and empathic concern. *Frontiers in neuroscience, 7*.

Mesoudi, A. (2016). Cultural evolution: integrating psychology, evolution and culture. *Current Opinion in Psychology, 7*, 17-22.

Meyer-Lindenberg, A., Domes, G., Kirsch, P., & Heinrichs, M. (2011). Oxytocin and vasopressin in the human brain: social neuropeptides for translational medicine. *Nature Reviews Neuroscience, 12*(9), 524-538.

Neumann, I. D., Maloumby, R., Beiderbeck, D. I., Lukas, M., & Landgraf, R. (2013). Increased brain and plasma oxytocin after nasal and peripheral administration in rats and mice. *Psychoneuroendocrinology, 38*(10), 1985-1993.

Nikiforakis, N., & Mitchell, H. (2014). Mixing the carrots with the sticks: third party punishment and reward. *Experimental Economics, 17*(1), 1-23.

Paloyelis, Y., Doyle, O. M., Zelaya, F. O., Maltezos, S., Williams, S. C., Fotopoulou, A., & Howard, M. A. (2014). A spatiotemporal profile of in vivo cerebral blood flow changes following intranasal oxytocin in humans. *Biological Psychiatry, 79*(8), 693-705.

Rimmele, U., Hediger, K., Heinrichs, M., & Klaver, P. (2009). Oxytocin makes a face in memory familiar.

The Journal of Neuroscience, 29(1), 38-42.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*: CUP Archive.

Sakamoto, H., Matsuda, K.-i., Hosokawa, K., Nishi, M., Morris, J. F., Prossnitz, E. R., & Kawata, M. (2007). Expression of G protein-coupled receptor-30, a G protein-coupled membrane estrogen receptor, in oxytocin neurons of the rat paraventricular and supraoptic nuclei. *Endocrinology, 148*(12), 5842-5850.

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior, 35*(3), 169-175.

Shalvi, S., & De Dreu, C. K. W. (2014). Oxytocin promotes group-serving dishonesty. *Proceedings of the National Academy of Sciences, 111*(15), 5503-5507.

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies:"Altruistic" punishment of in-group members. *Evolution and Human Behavior, 25*(6), 379-393.

Striepens, N., Kendrick, K. M., Hanking, V., Landgraf, R., Wüllner, U., Maier, W., & Hurlemann, R. (2013). Elevated cerebrospinal fluid and blood concentrations of oxytocin following its intranasal administration in humans. *Scientific reports, 3*.

Tatsuoka, M. M., & Lohnes, P. R. (1988). *Multivariate analysis: Techniques for educational and psychological research*: Macmillan Publishing Co, Inc.

Ten Velden, F. S., Baas, M., Shalvi, S., Kret, M. E., & De Dreu, C. K. (2014). Oxytocin differentially modulates compromise and competitive approach but not withdrawal to antagonists from own vs. rivaling other groups. *Brain research, 1580*, 172-179. doi: 10.1016/j.brainres.2013.09.013

Ten Velden, F. S., Daughters, K., & De Dreu, C. K. W. (2016). Oxytocin promotes intuitive rather than deliberated cooperation with the in-group. *Hormones and Behavior*.

Tyler, T. R., & Fagan, J. (2008). Legitimacy and cooperation: Why do people help the police fight crime in their communities? *Ohio St. J. Crim. L., 6*, 231.

Van IJzendoorn, M. H., Bhandari, R., Van der Veen, R., Grewen, K. M., & Bakermans-Kranenburg, M. J. (2012). Elevated salivary levels of oxytocin persist more than 7 h after intranasal administration. *Frontiers in neuroscience, 6*.

Weisman, O., Zagoory-Sharon, O., & Feldman, R. (2012). Intranasal oxytocin administration is reflected in human saliva. *Psychoneuroendocrinology, 37*(9), 1582-1586. doi: 10.1016/j.psyneuen.2012.02.014

Winer, B. J., Brown, D. R., & Michels, K. M. (1971). *Statistical principles in experimental design* (Vol. 2). New York: McGraw-Hill

Figure Legends

Figure 1 Timeline of the Experimental Procedure and the Third-Party Punishment and Reward Trust Game (TPPR-TG).

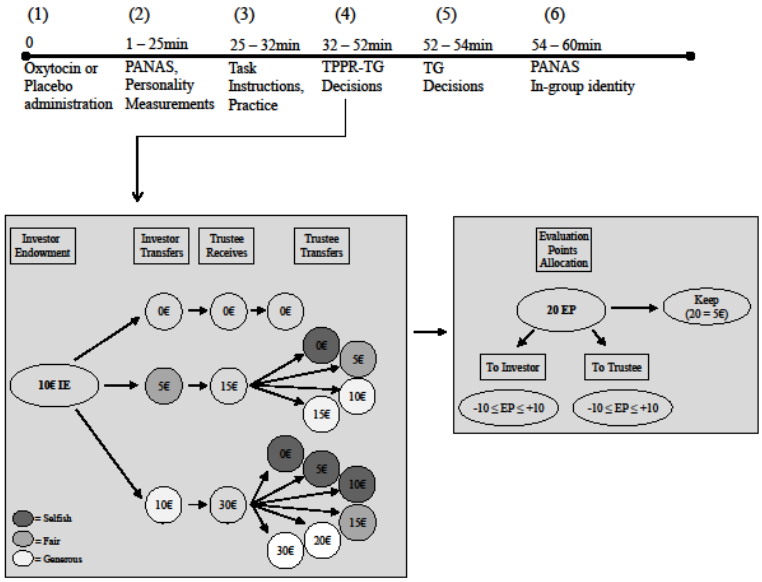

Figure 2 Mean allocations to investors as a function of investor group and treatment (range -10 to +10; displayed Mean ± SE). Connectors indicate $p<0.01$.

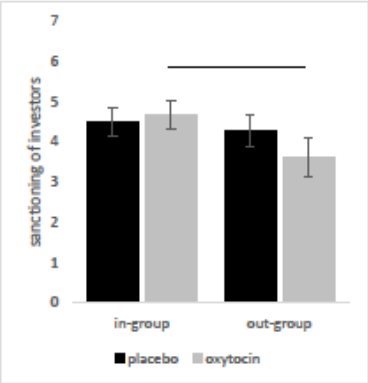

Figure 3 Oxytocin modulates sanctioning of in-group and outgroup investors. **(A)** Sanctioning of investors when trustees' back-transfers were selfish (range -10 to +10; displayed Mean ± SE). **(B)** Sanctioning of investors when trustees' back-transfers were fair (range -10 to +10; displayed Mean ± SE). **(C)** Sanctioning of investors when trustees' back-transfers were generous (range -10 to +10; displayed Mean ± SE). Connectors indicate $p<0.05$ (Bonferroni corrected).

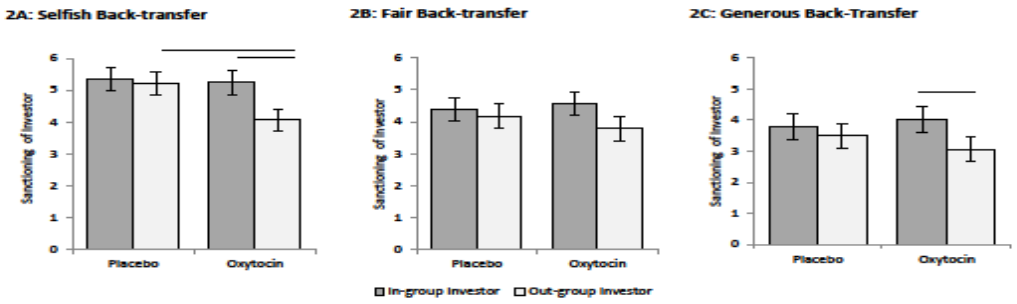

Figure 4 Oxytocin modulates sanctioning of in-group and outgroup trustees. **(A)** Sanctioning of trustees when trustees' back-transfers were selfish (range -10 to +10; displayed Mean ± SE). **(B)** Sanctioning of trustees when trustees' back-transfers were fair (range -10 to +10; displayed Mean ± SE). **(C)** Sanctioning of trustees when trustees' back-transfers were generous (range -10 to +10; displayed Mean ± SE). Connectors indicate $p<0.05$ (Bonferroni corrected).