OXFORD

Social Cognitive and Affective Neuroscience, 2017, 1–8

doi: 10.1093/scan/nsx101
Original Article

Oxytocin promotes altruistic punishment

Gökhan Aydogan,[1] Nadja C. Furtner,[2] Bianca Kern,[2] Andrea Jobst,[3] Norbert Müller,[3,4] and Martin G. Kocher[2,5,6]

[1]Department of Psychology, Arizona State University, Tempe, AZ 85281, USA, [2]Department of Economics, Ludwig-Maximilians University Munich, Germany, [3]Department of Psychiatry and Psychotherapy, Ludwig-Maximilians University Munich, Germany, [4]Marion von Tessin Memory Center, Munich, Germany, [5]Institute for Advanced Studies, Vienna, 1080 Vienna, Austria, and [6]Department of Economics, University of Gothenburg, Gothenburg, Sweden

Correspondence should be addressed to Gökhan Aydogan Department of Psychology, Arizona State University, 950 S. McAllister Ave., Tempe, AZ 85287. E-mail: goekhan.aydogan@asu.edu.

Abstract

The role of neuromodulators in the enforcement of cooperation is still not well understood. Here, we provide evidence that intranasal applied oxytocin, an important hormone for modulating social behavior, enhances the inclination to sanction free-riders in a social dilemma situation. Contrary to the notion of oxytocin being a pro-social hormone, we found that participants treated with oxytocin exhibited an amplification of self-reported negative social emotions such as anger towards free-riders, ultimately resulting in higher magnitude and frequency of punishment of free-riders compared to placebo. Furthermore, we found initial evidence that oxytocin contributes to the positive effects of a punishment institution by rendering cooperation preferable in the oxytocin condition for even the most selfish players when punishment was available. Together, these findings imply that the neural circuits underlying altruistic punishment are partly targeted by the oxytonergic system and highlight the importance of neuromodulators in group cohesion and norm enforcement within social groups.

Key words: oxytocin; neuroendocrinology; social dilemma; altruistic punishment; norm enforcement

Introduction

Oxytocin constitutes one of the most important neuromodulators of social behavior among mammals, including humans. The evidence on its exact mechanisms of action, however, is still inconclusive. Several studies have found that the neuropeptide oxytocin modulates various behaviors associated with pro-social behavior (Kosfeld *et al.*, 2005; Israel *et al.*, 2009; Mikolajczak *et al.*, 2010; Israel *et al.*, 2012), including conflict resolution (Ditzen *et al.*, 2009), in-group conformity (Stallen *et al.*, 2012) and both cognitive and emotional empathy (Domes *et al.*, 2007; Rodrigues *et al.*, 2009, Guastella *et al.*, 2010; Schulze *et al.*, 2011; Shamay-Tsoory *et al.*, 2013). It is these characteristics that have led to the common interpretation of oxytocin (henceforth OT) as a 'pro-social' hormone (MacDonald and MacDonald, 2010, Ebitz *et al.*, 2013). This notion, however, has recently been questioned due to contradictory findings regarding the effects of inhaled OT on pro-social preferences. For instance, a recent systematic review of the oxytocin literature found no evidence that trust behavior is associated with oxytocin (Nave *et al.*, 2015). Specifically, these authors found that trust was neither influenced by inhaled OT, nor by OT plasma levels nor by any genetic polymorphisms of the OT receptor gene.

Moreover, several recent studies indicate a more contextual effect of OT on social behavior, since OT has been shown to promote cooperation within groups but not between them (Dreu *et al.*, 2010), to enhance a general ethnocentric bias (De Dreu *et al.*, 2011), to increase dishonesty (Shalvi and De Dreu, 2014; Aydogan *et al.*, 2017), and to amplify negative social emotions like envy and schadenfreude evoked by unfair money allocations (Shamay-

Received: 19 January 2017; **Revised:** 2 August 2017; **Accepted:** 17 August 2017

Tsoory *et al*., 2009). Additionally, inhaled OT increases sensitivity to the social information of co-players in a social dilemma game (Declerck *et al*., 2010, 2013; Mikolajczak *et al*., 2010; Bartz *et al*., 2011;), indicating a more complex role of the neuropeptide in cooperative behavior (Bartz *et al*., 2011) than has often previously been assumed. This more differentiated view is also supported by animal studies showing that endogenous OT release in the rodent brain correlates with aggression in mate-guarding behavior (Bales and Carter, 2003) and with maternal aggression in defending offspring against intruders (Bosch, 2005; Bosch and Neumann, 2012).

In this work, we examine the effect of OT on the enforcement of cooperation, since it has been shown that the inclination to altruistically punish uncooperative behavior with no egoistic material benefits is a crucial factor for the sustainability of cooperation (Fehr and Schmidt, 1999; Fehr and Gächter, 2002; Fehr and Fischbacher, 2004). In building upon studies suggesting that OT is associated with social emotions like envy and schadenfreude in unfair money allocations (Shamay-Tsoory *et al*., 2009), it is straightforward to assume that inhaled OT may amplify negative emotions (e.g. 'anger') toward defectors and therefore increase the inclination to enforce the norm of cooperation.

Our primary hypothesis concerns punishment inclination in a social dilemma game. We therefore hypothesize that inhaled OT increases the inclination to punish defectors relative to a placebo in a social dilemma game. To confirm previous findings on the motivational factors of altruistic punishment (Hopfensitz and Reuben, 2009), we also asked participants to indicate the emotions they experienced before and after the social dilemma game. As our supporting hypothesis, we predicted an amplification of negative emotions directed toward defectors following the violation of a social norm.

Finally, to explore OT's effects on the efficiency of a punishment institution, we also conducted further analyses examining the directionality and the effectiveness of punishment—important features that may determine the sustainability of the social norm.

Materials and methods

Design and participants

To examine the effect of oxytocin on altruistic punishment in a social dilemma situation, we employed a double-blind, randomized, placebo-controlled between-subjects design, in which participants intra-nasally self-administered oxytocin ($n = 72$) or a placebo ($n = 72$) (Syntocinon-Spray, Sigma-Tau; 3 puffs per nostril, each with 4 IU of oxytocin).

A recent meta-analysis by Valstad *et al*. (2017) confirmed that inhaled OT constitutes a reliable and harmless way to manipulate participants' OT brain levels. However, several recent reviews identified relatively poor replicability within the OT literature (Nave *et al*., 2015; Leng and Ludwig, 2016; Walum *et al*., 2016), which partly arises from small sample sizes. To account for these statistical issues, we recruited 144 male participants—a substantially higher sample size compared to similar studies that, according to Walum *et al*., exhibit a median of 49 individuals.

Participants were between 18 and 33 years old ($M \pm SD$, 23.7 ± 3.1 years) and were recruited via ORSEE (Greiner, 2015). Exclusion criteria for participation were significant medical or psychiatric disorders, medication, smoking >15 cigarettes per day and drug or alcohol abuse. Participants were instructed to abstain from food for 1 h before the experiment and from alcohol, smoking and caffeine for at least 24 h. We also informed all participants during the recruitment that this experiment's goal is to study the influence of oxytocin on economic behavior. The study was approved by the ethics committee of the Department of Medicine at the University of Munich and written informed consent was obtained by all participants before participation. Analyses were conducted on all 144 participants. The whole experiment took <90 min and was programmed using z-Tree (Fischbacher, 2007). All games involved real monetary stakes (denoted MU for 'monetary units' in the following and converted to euro amounts after the experiment). Each MU earned in the experiment was worth 0.2 € and in the end of the experiment, all participants received a flat payment of 4 € as well as their payoff corresponding to their success in the experiment (mean payoff ± s.d., 28.34 ± 10.4 €).

Experimental procedure

Participants received neutrally written instructions (see Instructions in the Supplementary Material), which we read aloud to make them common knowledge. Decisions were made anonymously on PC screens, and participants were separated into cubicles.

Fifty minutes after the administration of the nasal spray, the relevant part of the incentivized experiment started. In each session, participants were asked to play a one-shot prisoners' dilemma game without a punishment opportunity (henceforth, PDGX) and with a punishment (sanctioning) opportunity (henceforth, PDGS) (Falk *et al*., 2005). In both treatments PDGX and PDGS, we made sure, with control questions, that participants understood the payoff structure. To exclude strategic or reputational concerns, participants were randomly and anonymously assigned to groups of three in both treatments. They were informed that they would not be matched with the same people in each of the two games or in any other part of the experiment.

In PDGX, participants had to decide simultaneously in groups of three whether to cooperate or defect. The incentives of PDGX game are outlined in Table 1. Feedback on the other group members' decisions and on the payoff from PDGX was only given at the end of the entire experiment in order to avoid learning or spillover effects. In game PDGX participants were asked to make the single choice either to cooperate or to defect. As can be seen, defecting is the payoff-maximizing and hence dominant strategy for each player, independent of the actions of the other players in this group. Collectively, however, the highest and thus welfare-maximizing outcome for the group can only be reached through cooperation, fulfilling the criteria for a social dilemma.

Then the game PDGS started, which consisted of two stages. In the first stage, participants played a three-person prisoners' dilemma game equivalent to PDGX. In the second stage, participants had the opportunity to sanction after being informed about the decisions (cooperate/defect) of their group members in stage 1. They were informed about this opportunity and about the fact that their own decisions would be made visible to their group members as well at the beginning of the treatment, before all decisions were made. All participants could assign up to a maximum of 10 punishment points to each of the other two members in their group. These punishment points were neutrally denoted 'deduction points' in the experimental instructions. Each punishment point assigned, for which the recipient was deducted 3 MU, cost the punishing participant 1 MU. The final individual payoff π_i for individual i in the PDGS condition was thus determined by the income from the first stage, minus three times the sum of received punishment points, minus the sum of assigned punishment points:

$$\pi_i = \max\left[0,\ 20 - g_i + 0.6\sum_j g_j - 3\sum_{j\neq i} p_{ji}\right] - \sum_{j\neq i} p_{ij},$$

where g_i denotes the contribution of subject i, the sum of g_j denotes the contributions of all three group members to the public

Table 1. Payoffs to player *i* in PDGX and the first stage of PDGS (Falk *et al.*, 2005)

	Both other players defect	One of the other two players cooperates	Both other players cooperate
Player *i* defects	20	32	44
Player *i* cooperates	12	24	36

Notes and Sources: Participants were randomly and anonymously assigned to groups of three. They decided simultaneously whether to cooperate or to defect. For example, if player *i* decided to defect whereas both other group members decided to cooperate, player *i* would earn an income of 44 monetary units (MU) in PDGX and/or in the first stage of PDGS. If instead player *i* decided to cooperate, she would earn an income of 36 MU.

good, p_{ij} represents the amount of deduction points subject *i* assigns to subject *j* with $i \neq j$, and p_{ji} depicts the amount of deduction points subject *i* receives by subject *j*. Since the first stage of PDGS is equivalent to the PDGX, the choice to contribute is binary such that *g* is defined as $g \in \{0,\ 20\},\ \forall\ i,j$. As punishment is costly, the dominant strategy for selfish decision makers is not to punish (see Supplementary Material for proof).

To measure their emotional states, participants rated 8 emotions on a 7-point Likert scale, which was conducted twice: first, immediately after substance administration and, second, immediately after the norm enforcement decision in PDGS. Participants had to assess the valence of positive and negative emotions, namely anger, gratitude, guilt, joy, irritation, shame, surprise and disappointment (Hopfensitz and Reuben, 2009). We also asked participants to rate the perceived fairness of the other group members' decision in the first stage in the post-experimental questionnaire. Participants received an additional 20 MU for the completion of the questionnaires.

Results

To examine whether OT affects the likelihood and magnitude of punishment directed towards defectors, we ran a Tobit regression (see Table 2) that controlled for the overall cooperation level within a group and the directionality of punishment incidents. Specifically, we regressed the received punishment points of subject *i* on his or her decision to negatively or positively deviate from the other group member *j*'s decision to cooperate, the group's overall level of cooperation, the specific condition (OT *vs* placebo), and the interaction between defection and OT treatment. The first regression in Model 1 reveals that participants receive deduction points if they negatively deviate, i.e. if they defect when the other group member *j* chooses to cooperate, but not if they positively deviate (i.e. if *i* cooperates but *j* does not). Furthermore, others' cooperation level has no impact on being punished. Model 2 extends Model 1 by OT and its interaction with free-riding behavior, and shows that a negative deviation leads to a harsher punishment in the OT treatment compared to the placebo treatment. Again, OT alone has no significant impact on the amount of punishment points a subject receives. It is rather the interaction between OT and a negative deviation that leads to a significantly higher amount of received deduction points.

This result is confirmed by the finding that punishment of cooperators was virtually absent in both treatment groups (Figure 1B), without any statistical difference between them (Mann–Whitney *U*-test; $z = -0.578$, $P = 0.5630$, two-sided). Conversely, we find that defectors in the OT group received an average of 4.3 punishment points (Figure 1A), which is almost twice as much as the amount of 2.2 punishment points imposed on defectors in the placebo group (Mann–Whitney *U*-test; $z = 2.334$, $P = 0.0196$, two-sided). Since one deduction point in this experiment corresponds to a payoff reduction of 3 MU, a punished defector would face an average reduction of his payoff by 12.9 MU ($SD = 11.39$) in the OT condition compared to a reduction of only 6.6 MU ($SD = 11.12$) in the placebo condition. We therefore conclude that not only is the inclination to cooperate enhanced with OT, but the degree of punishment is increased as well, indicating that OT positively influences the credibility of using altruistic punishment to enforce cooperation.

Moreover, to examine the effect of OT on punishment frequency, we analyzed the directionality of each punishment incident. Figure 2 displays the direction of all punishment activity and confirms that the vast majority of punishment incidents go from cooperators to defectors. This result shows that OT does not trigger a stronger inclination to punish others, regardless of the violation of norms: for instance, because of undirected emotional arousal or spiteful motives. Such reasoning would imply that the punishment behavior of defectors against both, cooperators and other defectors, differs between the OT group and the placebo group. As can be seen in Figure 2, this is not the case. Punishment incidents by defectors are not more frequent in the OT group than in the placebo group (4 out of 44 in the OT group *vs* 6 out of 68 in the placebo group; Fisher's exact-test; $P = 1$, two-sided). However, there is a difference in the inclination to punish defectors: In the placebo group, roughly 22% (15 out of 68) of all defectors were punished by cooperators. By contrast, about 43% (19 out of 44) of defectors in the OT group were punished, which is significantly higher than the punishment rate in the placebo group ($\chi^2 = 5.6379$; $df = 1$; $P = 0.018$, two-sided). Thus, we conclude that OT exhibits both, an amplification effect on the punishment magnitude as well as on the punishment rate toward defectors.

However, the question still remains whether the motive for punishment in the OT group is indeed fairness driven or possibly rooted in other motives. Therefore, we analyzed the modulation of negative emotions, triggered by the violation of a social norm. Specifically, negative emotions related to the violation of fairness norms, such as anger and disappointment, should be intensified under the influence of OT and thus should trigger punishment acts (Shamay-Tsoory *et al.*, 2009). To test this hypothesis, participants were asked to report the intensity of several positive and negative emotions on a seven-point Likert scale immediately after the allocation of their punishment points and—in order to control for baseline effects—at the beginning of the experiment, subsequent to substance administration. As is typical in survey questions on emotions, a list of related and probably unrelated emotions is presented in order to avoid experimenter demand effects in elicitation (the focus here was on anger and disappointment; details on the rating of other emotions can be found in the Supplementary Material). Our data shows that participants given OT reported higher levels of anger towards defectors ($M = 3.07$ *vs* 1.65; $SD = 2.59$ *vs* 2.57; Mann–Whitney *U*-test; $z = -2.701$, $P = 0.007$, two-sided, controlled for baseline emotions) and higher levels of disappointment ($M = 2.95$ *vs* 2.20; $SD = 2.47$ *vs* 2.23; Mann–Whitney *U*-test;

Table 2. Influences on received deduction points

	Dependent variable: Received deduction points of participant i	
Ind. Variables	Model (1)	Model (2)
Cooperation level in group of *i*	0.266 (1.148)	−0.604 (1.293)
Negative deviation of *i*	10.104*** (1.476)	7.329*** (1.779)
Positive deviation of *i*	−0.285 (1.867)	−3.019 (3.230)
OT × Negative deviation of *i*	–	5.707** (2.474)
OT × Positive deviation of *i*	–	4.175 (3.356)
OT	–	−1.877 (2.146)
Constant	−8.308*** (2.289)	−6.133** (2.484)
R^2	0.100	0.108

Notes and Sources: Note that ***, **, * denote significance at the 1%, 5% and 10% level. Following Fehr and Gächter (2000), we performed a Tobit regression with the number of received punishment points of *i* as the dependent variable ($N = 288$) and clustered standard errors on the participant level. Because the number of received deduction points ranged from 0 to 20, we used a Tobit regression to account for the censored nature of the dependent variable. The overall level of cooperation in *i* 's group is defined as 0 if both other group members (other than *i*) defect, 1 if one of them cooperates and 2 if both other group members choose to cooperate. The variable negative deviation is 1 if subject *i* defects and player *j* cooperates or 0 otherwise. The variable positive deviation is 1 if subject *i* cooperates and player *j* defects or 0 otherwise. Additionally, in model 2 we calculated the interaction effect of OT with negative and positive deviation of *i*. OT was a dummy variable for the oxytocin group (=1) and the placebo group (=0). An OLS regression provides similar results.

$z = -1.800$, $P = 0.0718$, two-sided, controlled for baseline emotions). Furthermore, a post-experimental survey question reveals that participants in the OT group also perceived the behavior of defecting group members as significantly less fair than participants in the placebo group ($M = 2.93$ *vs* 3.39; $SD = 2.09$ *vs* 1.66; Mann–Whitney *U*-test; $z = 1.860$, $P = 0.0629$, two-sided). This indicates that participants given OT were emotionally more prone to anger in the face of violated social norms and suggests a significant impact of OT on neural circuits encoding emotional responses in social interactions.

Our results finally raise the question whether the punishment magnitude is sufficient to constitute a credible threat against defective behavior. We therefore computed the expected payoff given the decisions of all group members in the first and second stage of the PDGS. Thus, we computed the average payoffs given each possible decision in the first stage and then subtracted the average punishment corresponding to the respective action set with:

$$E(\pi_i|g_i, g_j) = 20 - g_i + 0.6\sum_j g_j - E\left(3\sum_{j\neq i} p_{ji}(g_i, g_j)\right)$$

In Table 3, the average payoff of subject *i* is illustrated as a function of the other group members' decisions in the oxytocin and the placebo condition.

Based on the payoff matrices in Table 3 and the frequency of all players' actions, we were able to compute the expected

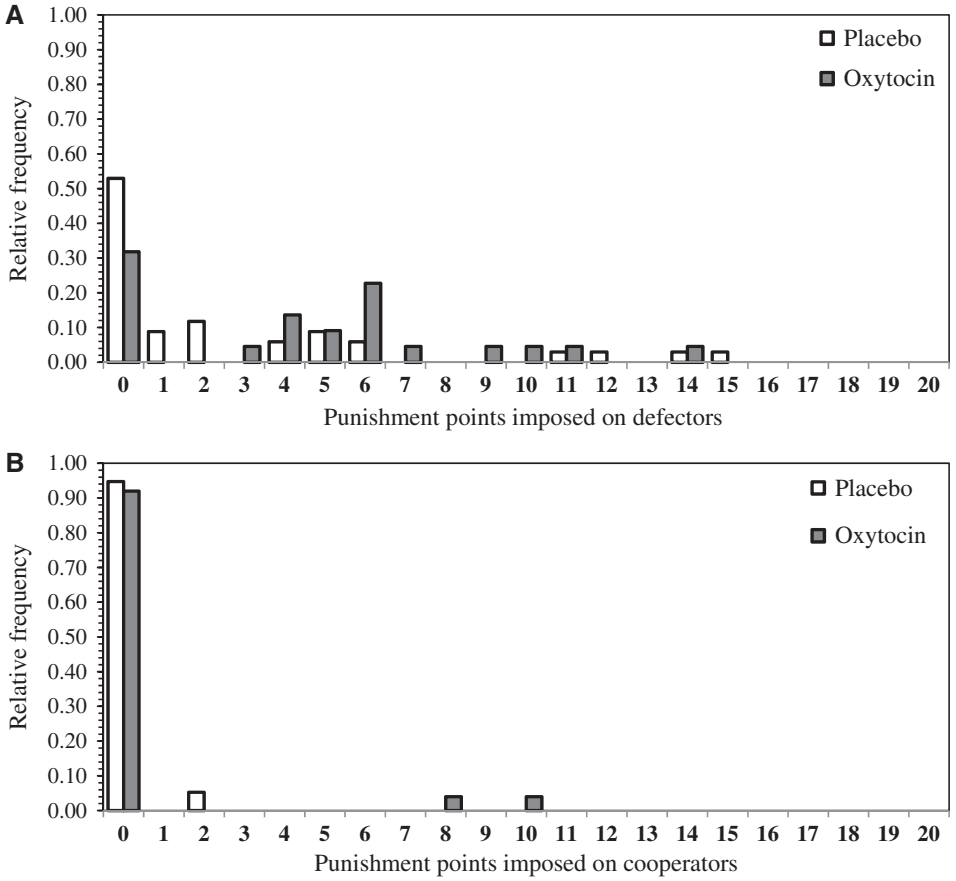


Fig. 1. Relative frequency of assigned punishment points imposed on defectors (A) and on cooperators (B) as a function of oxytocin. Assigned punishment points are depicted for the OT group in gray bars and for the placebo group in white bars ($n = 144$). (**A**) Punishment points imposed on defectors ($M = 4.32$, $SD = 3.79$) are significantly higher in the OT group than on defectors ($M = 2.20$, $SD = 3.70$) in the placebo group (Mann–Whitney *U*-test; $z = -2.334$, $P = 0.0196$, two-sided). (**B**) Punishment of cooperators is almost nonexistent; subjects in the OT group ($M = 0.72$, $SD = 2.48$) and the placebo group ($M = 0.10$, $SD = 0.45$) show non-distinguishable inclinations to punish cooperators (Mann–Whitney *U*-test; $z = -0.578$, $P = 0.5630$, two-sided).

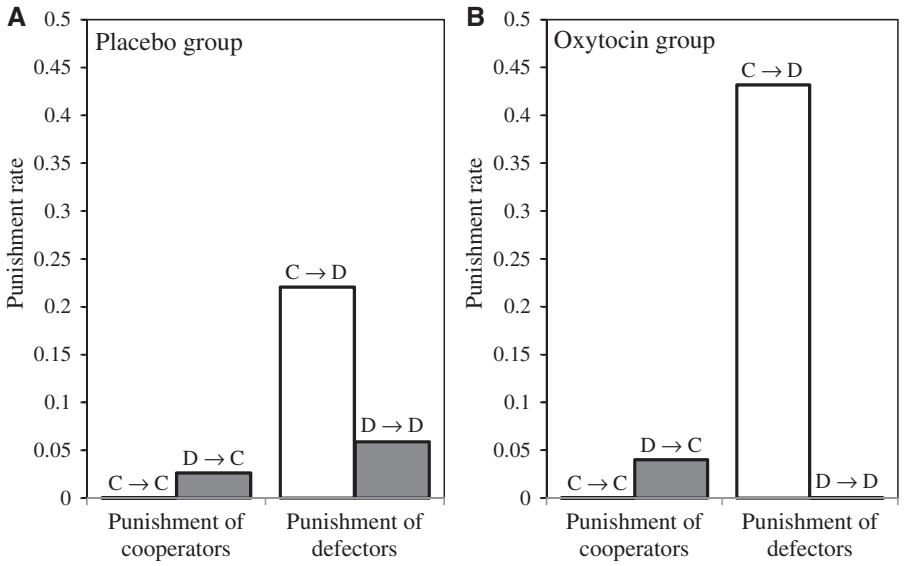


Fig. 2. Punishment rate by punished types. The bar C→C represents the proportion of all cooperators punished by other cooperators; analogously, C→D is the proportion of all defectors punished by cooperators; and so on. (**A**) In the placebo group about 22% (15 out of 68) of all defectors were punished by cooperators, whereas only 6% (4 out 68) were punished by other defectors. Punishment of cooperators (2 out of 76) played only a minor role. (**B**) In the OT group 43% (19 out of 44) of all defectors were punished by cooperators. A few defectors punished cooperators (4 out of 100).

Table 3. Payoff of player *i* in PDGS including all expected punishments of the second stage for the oxytocin (A) and the placebo (B) condition

		Both other players defect	One of the other two players cooperates	Both other players cooperate
(A) Oxytocin	Player *i* defects	20 (0)	27.5 (8.33)	26.21 (10.15)
	Player *i* cooperates	12 (0)	20.14 (9.69)	36 (0)
(B) Placebo	Player *i* defects	20 (0)	26.46 (9.9)	27.8 (15.5)
	Player *i* cooperates	11.07 (2.2)	24 (0)	36 (0)

Notes and Sources: The corresponding payoffs in the first stage are reduced by the average punishment given the decisions of all group members in the first stage. Standard deviations are depicted in parentheses. Due to altruistic punishment, playing 'defect' is neither in the oxytocin (A) nor in the placebo (B) condition a dominant strategy.

payoff of a participant by utilizing the actual distribution of cooperators and defectors in our sample. Consequently, in the placebo condition free riding would still result in a higher expected payoff, since the expected payoff of defecting with 25.39 MU is higher than the expected payoff of cooperating with 24.46 MU. Assuming rational expectations, a rational and selfish agent would therefore prefer to defect in the placebo condition. Conversely, participants in the OT group are better off cooperating, since cooperation leads to an expected payoff of 26.74 MU compared to 26.51 MU in case of defecting. So, for participants in the OT condition it would be unprofitable to defect in the first stage, given the expected payoff reduction by other players in the second stage. Consequently, a rational and selfish decision-maker would prefer cooperating in the OT but not in the placebo condition.

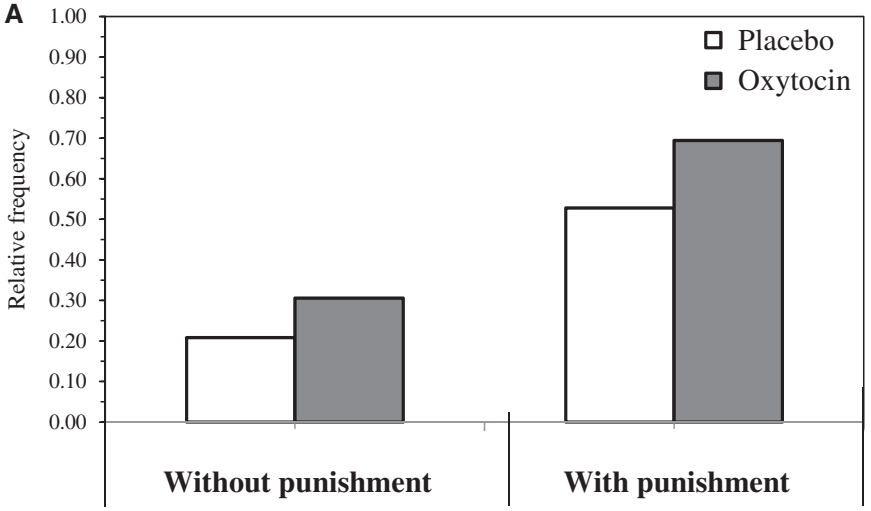
To test whether the credible threat of punishment was indeed sufficient to achieve the observed increase in cooperation, we analyzed cooperation behavior separately with and without the punishment option (see Figure 3A). To analyze both games, we ran two different Probit regressions to account for (i) positive effects of a sanctioning mechanism on cooperation, (ii) general positive effects of OT on cooperation and (iii) an interaction effect of both variables. The first regression model in Figure 3B reveals that both punishment (Model 1: $b_{PO} = 0.359$, $P < 0.001$) and OT (Model 1: $b_{OT} = 0.146$, $P = 0.035$) have a significantly positive impact on cooperation. However, to account for possible interaction effects between both conditions, we ran a second Probit regression (see Model 2). We find that the punishment option shows a significant influence on cooperation (Model 2: $b_{PO} = 0.335$, $P < 0.001$), whereas the interaction effect fails to reach significance ($b_{OTxPO} = 0.053$, $P = 0.602$).

Our results suggest that, while cooperation levels are significantly higher when punishment of free-riders is possible compared to when it is not—a result that has been established multiple times in the literature (Fehr and Gächter, 2000, 2002; Falk *et al.*, 2005; Chaudhuri, 2011)—OT also shows a positive effect on cooperation, irrespective of the availability of a sanctioning mechanism. This implies that OT contributes to the efficiency of sanctioning mechanisms predominantly by increasing the inclination to punish, and indirectly by increasing the likelihood of cooperation.

Discussion

The present study provides initial evidence for the role of OT in the enforcement of cooperation within small groups. In particular, we show that inhaled OT significantly increases both the

A



B

	Dep. Variable: *Choice to cooperate* (*defect = 0; cooperate = 1*)	
Ind. Variables	(1)	(2)
Punishment Option	.359*** (.045)	.335*** (.067)
Oxytocin	.146** (.069)	.118 (.088)
Oxytocin×Punishment Option		.053 (.101)
Pseudo R^2	.1103	.1108

Fig. 3. (**A**) Relative frequency of cooperation levels with and without a punishment option in the OT (grey bars) and placebo (white bars) groups ($n = 144$). (**B**) The choice to cooperate is the dependent variable in a Probit regression with clustered standard errors on the participant level. Coefficients represent marginal effects on the probability to cooperate. Note that ***, **, * denote significance at the 1%, 5% and 10% level.

likelihood and magnitude of punishment of uncooperative behavior compared to a placebo, and this increase is accompanied by an amplification of negative emotional reactions toward defectors. Furthermore, our data support the notion that OT generally promotes cooperation, irrespective of whether a punishment option is present. Our results additionally imply that inhaled OT contributes to the efficiency of sanctioning mechanisms directly by increasing the inclination to punish uncooperative behavior and indirectly by increasing the overall likelihood of cooperation.

According to previous findings, OT is thought to make norm adherence more likely through a positive impact on positive emotions, resulting in an enhanced pro-social attitude (Kosfeld *et al.*, 2005; Ditzen *et al.*, 2009; Andari *et al.*, 2010; De Dreu *et al.*, 2010; MacDonald and MacDonald, 2010; Mikolajczak *et al.*, 2010; Bartz *et al.*, 2011; Israel *et al.*, 2012; Rilling *et al.*, 2012; Declerck *et al.*, 2013). However, our results suggest a different perspective on the popularly known 'moral molecule', as we show that OT, rather than having an effect on positive emotions, amplifies strong negative emotions (i.e. anger) towards non-cooperators within small groups. This remarkably strong emotional reaction leads to harsher punishment of uncooperative behavior in the OT group compared to the placebo group. Consistent with the finding that emotions, in addition to rational considerations, work as a proximate mechanism to induce norm-enforcing behavior (Fehr and Gächter, 2002; De Quervain *et al.*, 2004; Falk *et al.*, 2005; Hopfensitz and Reuben, 2009), our data suggests that OT might have an amplifying effect on social emotions, including negative, which ultimately triggers the punishment of defective behavior and leads to the enforcement of social norms. This notion is supported by recent findings, indicating that OT can increase other emotions considered negative, like envy or gloating (Shamay-Tsoory *et al.*, 2009). Reflecting this evidence, more research is needed to indicate whether OT has a general amplifying effect on emotions in social interactions, which would align with the observed context-dependent role of OT in social behavior (Bartz *et al.*, 2011).

Furthermore, our data indicate that, regardless of the availability of a punishment option, cooperation rates are significantly higher following the administration of OT relative to a placebo. This general increase in the inclination to cooperate may indirectly contribute to the efficiency of a sanctioning mechanism. That is, since fewer sanctioning incidents would be required to achieve the same level of cooperation within a group, higher cooperation rates would render the total costs of norm enforcement significantly lower and therefore lead to higher aggregated payoffs within the group.

This notion is also in line with the actual punishment behavior in our data. Given the punishment behavior in both conditions, participants in the OT condition would always be better off cooperating, whereas participants in the placebo condition would still have the opportunity to increase their payoff by free-riding (as the magnitude of punishment is not sufficiently high to render cooperation profit-maximizing). Consequently, a selfish agent with rational expectations would be better off defecting in the placebo condition and better off cooperating in the OT condition.

We therefore conclude that participants treated with OT show a higher inclination to deter uncooperative behavior,

which transforms the cooperation incentives from a classical prisoner's dilemma into a coordination game (Fehr and Schmidt, 1999; Fehr and Gächter, 2000, 2002; Fehr and Fischbacher, 2004). In a similar vein, Zak *et al.* (2007) found a positive effect of inhaled OT on monetary transfers in an ultimatum game, but not in a dictator game. The authors argue that this effect is rooted in the anticipation of putative punishment (through rejections of low offers)—a risk uniquely present in the ultimatum game. Given such beliefs, a rational decision-maker might therefore increase his or her expected payoff by reducing the risk of a rejected offer.

Some evidence also suggests that inhaled OT improves the accuracy of first order beliefs by enhancing one's ability to 'put oneself emotionally in the shoes of another person' (Domes *et al.*, 2007; Hollander *et al.*, 2007; Averbeck, 2010; Guastella *et al.*, 2010; Pedersen *et al.*, 2011; Schulze *et al.*, 2011; Shamay-Tsoory *et al.*, 2013). Here, we indirectly explored this by examining whether OT has an effect on the ability to form first order beliefs about other participants' inclination to cooperate or punish. Our data suggests no interaction effect between inhaled OT and the presence of a punishment option, which would be expected if inhaled OT increased the accuracy of first order beliefs. Nevertheless, more research is required to examine potential positive effects of OT on perspective taking in strategic settings.

Moreover, in this experiment we observe that the social norm of cooperation, which ties the group together, is defended aggressively in the prisoners' dilemma game against free-riders within the group. In a similar fashion, OT has been shown to stimulate defensive aggression (De Dreu *et al.*, 2010) against competing outgroups. Together, these results paint a broader picture, in which OT triggers aggressive behavior to protect the in-group's welfare, either against threats from outside or from within. This is also congruent with animal studies showing OT's role in triggering maternal aggression toward intruders or toward potential dangers to offspring (Bosch, 2005; Bosch and Neumann, 2012). The in-group (i.e. mother and children) is aggressively defended against potential threats to the group's well-being. Here, we show initial evidence in humans that OT can trigger or enhance anger, and subsequently result in harsher punishment of others. In our case, however, the threat to the group's well-being is not external; it arises from the actions of free-riders within the group itself. Whether the threat is internal or external, it is straightforward to assume that the oxytonergic system plays a crucial role in the modulation of behavior with the purpose to ultimately increase a social group's welfare, which is likely rooted in the ultimate goal to enhance the group's chances of survival. This notion is also supported by studies showing that OT enhances in-group conformity (Stallen *et al.*, 2012), in-group favoritism (De Dreu *et al.*, 2010, 2011) and group-serving dishonesty (Shalvi and De Dreu, 2014), all of which also serves group welfare.

Our results provide evidence for a deeper understanding of the oxytonergic system and give insights into its effects on cooperative behavior. Furthermore, our results call for more work on the connection of OT, social norms and norm-enforcement behavior, with the ultimate goal to understand the general underlying mechanism that leads to the various described effects of OT and to utilize this knowledge in potential clinical treatments of disorders associated to social behavior or norm adherence.

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

Author Contributions

G.A., N.C.F., B.K. and M.G.K. designed research; G.A., N.C.F. and A.J. performed research; G.A. analyzed data; G.A., N.C.F., B.K. drafted the manuscript, and M.G.K. provided critical revisions.

Funding

The nasal sprays containing oxytocin or placebo were prepared and delivered free of charge by Defiante Farmaceutica S.A. (Funchal, Portugal). The experiment was partly funded out of the regular university budget of the Chair in Behavioral and Experimental Economics, Ludwig-Maximilians-University Munich and the Foundation Immunität und Seele.

References

Andari, E., Duhamel, J.-R., Zalla, T., Herbrecht, E., Leboyer, M., Sirigu, A. (2010). Promoting social behavior with oxytocin in high-functioning autism spectrum disorders. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(9), 4389–94.

Averbeck, B.B. (2010). Oxytocin and the salience of social cues. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(20),9033–4.

Aydogan, G., Jobst, A., D'Ardenne, K., Müller, N., Kocher, M.G. (2017). The detrimental effects of oxytocin-induced conformity on dishonesty in competition. *Psychological Science*, **28**(6), 751–9.

Bales, K.L., Carter, C. (2003). Sex differences and developmental effects of oxytocin on aggression and social behavior in prairie voles (Microtus ochrogaster). *Special Issue on Aggressive and Violent Behavior*, **44**(3), 178–84.

Bartz, J.A., Zaki, J., Bolger, N., Ochsner, K.N. (2011). Social effects of oxytocin in humans: context and person matter. *Trends in Cognitive Sciences*, **15**(7), 301–9.

Bosch, O.J. (2005). Brain oxytocin correlates with maternal aggression: link to anxiety. *Journal of Neuroscience*, **25**(29), 6807–15.

Bosch, O.J., Neumann, I.D. (2012). Both oxytocin and vasopressin are mediators of maternal care and aggression in rodents: from central release to sites of action. *Hormones and Behavior*, **61**(3), 293–303.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, **14**(1), 47–83.

De Dreu, C.K.W., Greer, L.L., Handgraaf, M.J.J., *et al.* (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science*, **328**(5984), 1408–11.

De Dreu, C.K.W., Greer, L.L., Van Kleef, G.A., Shalvi, S., Handgraaf, M.J.J. (2011). Oxytocin promotes human ethnocentrism. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(4), 1262–6.

De Quervain, D.J.-F., Fischbacher, U., Treyer, V., *et al.* (2004). The neural basis of altruistic punishment. *Science (New York, N.Y.)*, **305**(5688), 1254–8.

Declerck, C., Boone, C., Kiyonari, T. (2013). The effect of oxytocin on cooperation in a prisoner's dilemma depends on the social context and a person's social value orientation. *Social Cognitive and Affective Neuroscience*, **9**(6), 802–9.

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), 368–74.

Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., Heinrichs, M. (2009). Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. *Biological Psychiatry*, **65**(9), 728–31.

Domes, G., Heinrichs, M., Michel, A., Berger, C., Herpertz, S.C. (2007). Oxytocin improves "mind-reading" in humans. *Biological Psychiatry*, **61**(6), 731–3.

Dreu, C.K.D., Greer, L.L., Handgraaf, M.J., *et al.* (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science*, **328**(5984), 1408–11.

Ebitz, R.B., Watson, K.K., Platt, M.L. (2013). Oxytocin blunts social vigilance in the rhesus macaque. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(28), 11630–5.

Falk, A., Fehr, E., Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, **73**(6), 2017–30.

Fehr, E., Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, **8**(4),185–90.

Fehr, E., Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, **90**(4), 980–94.

Fehr, E., Gächter, S. (2002). Altruistic punishment in humans. *Nature*, **415**(6868), 137–40.

Fehr, E., Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, **114**(3), 817–68.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, **10**(2), 171–8.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, **1**(1), 114–25.

Guastella, A.J., Einfeld, S.L., Gray, K.M., *et al.* (2010). Intranasal oxytocin improves emotion recognition for youth with autism spectrum disorders. *Biological Psychiatry*, **67**(7), 692–4.

Hollander, E., Bartz, J., Chaplin, W., *et al.* (2007). Oxytocin increases retention of social cognition in Autism. *Biological Psychiatry*, **61**(4), 498–503.

Hopfensitz, A., Reuben, E. (2009). The importance of emotions for the effectiveness of social punishment. *The Economic Journal*, **119**(540), 1534–59.

Israel, S., Lerer, E., Shalev, I., *et al.* (2009). The oxytocin receptor (OXTR) Contributes to prosocial fund allocations in the dictator game and the social value orientations task. *PLoS One*, **4**(5), e5535.

Israel, S., Weisel, O., Ebstein, R.P., Bornstein, G. (2012). Oxytocin, but not vasopressin, increases both parochial and universal altruism. *Psychoneuroendocrinology*, **37**(8), 1341–4.

Kosfeld, M., Heinrichs, M., Zak, P.J., Fischbacher, U., Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, **435**(7042), 673–6.

Leng, G., Ludwig, M. (2016). Intranasal oxytocin: myths and delusions. *Oxytocin and Psychiatry: From DNA to Social Behavior*, **79**(3), 243–50.

MacDonald, K., MacDonald, T.M. (2010). The Peptide That Binds. *Harvard Review of Psychiatry*, **18**(1), 1–21.

Mikolajczak, M., Gross, J.J., Lane, A., Corneille, O., Timary, P. d., Luminet, O. (2010). Oxytocin makes people trusting, not gullible. *Psychological Science*, **21**(8), 1072–4.

Nave, G., Camerer, C., McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspectives on Psychological Science*, **10**(6), 772–89.

Pedersen, C.A., Gibson, C.M., Rau, S.W., *et al.* (2011). Intranasal oxytocin reduces psychotic symptoms and improves Theory of Mind and social perception in schizophrenia. *Schizophrenia Research*, **132**(1), 50–3.

Rilling, J.K., DeMarco, A.C., Hackett, P.D., *et al.* (2012). Effects of intranasal oxytocin and vasopressin on cooperative behavior and associated brain activity in men. *Psychoneuroendocrinology*, **37**(4), 447–61.

Rodrigues, S.M., Saslow, L.R., Garcia, N., John, O.P., Keltner, D. (2009). Oxytocin receptor genetic variation relates to empathy and stress reactivity in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(50), 21437–41.

Schulze, L., Lischke, A., Greif, J., Herpertz, S.C., Heinrichs, M., Domes, G. (2011). Oxytocin increases recognition of masked emotional faces. *Psychoneuroendocrinology*, **36**(9), 1378–82.

Shalvi, S., De Dreu, C.K.W. (2014). Oxytocin promotes group-serving dishonesty. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(15), 5503–7.

Shamay-Tsoory, S.G., Abu-Akel, A., Palgi, S., *et al.* (2013). Giving peace a chance: oxytocin increases empathy to pain in the context of the Israeli–Palestinian conflict. *Psychoneuroendocrinology*, **38**(12), 3139–44.

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). *Biological Psychiatry*, **66**(9), 864–70.

Stallen, M., De Dreu, C.K.W., Shalvi, S., Smidts, A., Sanfey, A.G. (2012). The herding hormone: oxytocin stimulates in-group conformity. *Psychological Science*, **23**(11), 1288–92.

Valstad, M., Alvares, G.A., Egknud, M., *et al.* (2017). The correlation between central and peripheral oxytocin concentrations: a systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, **78**, 117–24.

Walum, H., Waldman, I.D., Young, L.J. (2016). Statistical and methodological considerations for the interpretation of intranasal oxytocin studies. *Oxytocin and Psychiatry: From DNA to Social Behavior*, **79**(3), 251–7.

Zak, P.J., Stanton, A.A., Ahmadi, S. (2007). Oxytocin increases generosity in humans. *PLoS One*, **2**(11), e1128.