Oxytocin makes us trusting but not gullible

Moïra Mikolajczak[1, 2]*, James J. Gross[3], Anthony Lane[1], Philippe de Timary[4] & Olivier Luminet[1, 2]

[1]*Université catholique de Louvain, Department of Psychology, Place Cardinal Mercier 10, B-1348 Louvain-La-Neuve, Belgium*

[2]*Belgian National Fund for Scientific Research (FNRS), Rue d'Egmont 5, B-1000 Brussels, Belgium*

[3]*Stanford University, Department of Psychology, Jordan Hall - Bldg 420, Stanford, CA 94305-2130, USA*

[4]*Université catholique de Louvain, Cliniques Universitaires Saint-Luc, Department of Neurology and Psychiatry, Avenue Hippocrate 10, B-1200 Brussels, Belgium*

Originally known for its role in childbirth and lactation, oxytocin (OT) has recently proved to play a key role in social behavior. Deprived of OT, humans are unable to recognize[1,2] and to bond[3] to their peers. Inversely, once boosted with OT, people become more caring[4], trusting[5,6,7] and generous[8]. Effect-sizes on trust and generosity were sufficiently large that OT started to be perceived as a natural drug that would make people credulous[9]. But could OT really impede judgment and lead individuals to trust untrustworthy peers? Here we show that oxytocin makes people trusting, but not gullible. Namely, OT did not have a trust-enhancing effect on people who interacted with seemingly unreliable peers. These results emphasize that the effect of OT is much more context-dependent than previously thought. This finding therefore invalidates some of the potential commercial or military applications of oxytocin[10].

"Danger breeds best on too much confidence" Pierre Corneille. Le Cid

In recent times, no hormone has fascinated the general public as much as oxytocin (OT). "A dose of human kindness", "love potion", and "liquid trust", are all nicknames that have been given to OT, a neuropeptide naturally secreted in the paraventricular and supraoptic nuclei. And these nicknames unquestionably have an element of truth: An increasing number of studies indicate that OT facilitates social relationships by altering both cognitions and behaviors in a pro-social way[5-8,11-14,]. For instance, oxytocin facilitates the encoding and memory of social stimuli[11,13], improves mind-reading[12], initates maternal behavior[4], and substantially increases trust[5-7] and generosity[8].

The effect of OT on generosity[8] and trust[5,6] was so large in previous studies (i.e., participants were 80% more trusting and generous in the OT group than in placebo group) that they unwittingly conveyed the idea that OT was a natural drug that made people nice and credulous[9]. This belief was then reinforced by a recent study showing that OT quickly restored trust after betrayals[6]. While the press and scientists[9] alike got concerned about potential misuses of these results by politicians, armed forces and marketers, OT retailers flourished—promising clients that spraying a thin mist of it over people would make them more trusting and willing to sign on the dotted line"[15].

But are these concerns warranted? Would OT really lead us to trust *anybody*? In previous studies on trust[5-7], participants never faced the same partner twice (i.e., they were never confronted again to a partner who breached their trust). Moreover, nothing pointed out that a specific partner could be unreliable. As a consequence, we do not know whether individuals in the OT group would have trusted an unreliable partner or maintained their trusting behavior in a betrayer. The aim of this study is to determine the circumstances under which OT increases trust: Does OT act independently of the

trustworthiness of one's peers (increasing our trust in both trustworthy and untrustworthy people) or does OT interact with the context (so that we become trusting only around trustworthy people)?

According to previous studies on animals, effects of OT are far from being linear or context-independent[16]. Namely, OT does not always lead individuals to behave pro-socially. Its effects seem instead to be determined by whether it is adaptive to be trusting in the given context (which largely depends on the trustworthiness of the target of the social interaction). In rodents, the female's OT release after giving birth makes her aggressiveness *lessened* towards her offspring, but *increased* towards potentially aggressive female intruders[17,18]. Thus, if OT is to facilitate both positive social interactions and survival, it should enhance trust but not credulity. Indeed, whereas trust contributes to economic and social success, credulity furthers market inefficiency[19] and social maladaptation[20]. We therefore speculated that OT would make people demonstrate more trust when interacting with trustworthy or neutral peers; conversely, we did not expect OT to have an effect on individuals who were interacting with untrustworthy peers.

We have tested our hypothesis in a customized version of the trust game[21,22] (see Figure 1). Sixty participants were either administered OT (experimental group) or a placebo (control group) and then played several rounds of the trust game with different partners, some seemingly more reliable than others. The trust game is a frequently used paradigm in neuroeconomics and behavioral economics, as it reproduces investors' trust dilemma in a lab. Each participant assumed the role of investor and could transfer money to a 'trustee', where it would triple. Subsequently, the trustee transferred all, a part, or none of the money back to the investor (see Methods section). If the investor entrusted the trustee with all of his money, the investor could maximize his profits if the trustee was reliable and fair. Conversely, he could lose everything if the trustee was not fair. The trust game is perfectly suited to

establish the investor's level of trust (i.e., the higher the trust, the higher the transfers). By manipulating the partners' trustworthiness, we have sought to determine to what extent OT impairs one's sensitivity to potential signs of dishonesty. Each participant played with three different types of trustees: seemingly reliable humans, seemingly unreliable humans, and the computer (i.e., fully neutral device). We hypothesized that investors from the OT group would transfer higher average amounts than those from the control group, unless there were hints that the trustee might not be trustworthy.

Analyses revealed that type of trustee strongly influenced the amount that participants invested in the trust game ($F = 65.44$, $p < .001$). On average, participants transferred less money to human partners than to the computer ($t_{59} = -5{,}75$, $p < .001$). As illustrated in table 1, this effect is largely due to unreliable trustees: Participants transferred significantly less money to unreliable trustees than to reliable ones ($p < .001$); reliable trustees, on the other hand, were transferred about as much money as the computer ($p = .145$).

Second, our data confirmed previous findings that OT substantially increases trusting behaviors ($F = 5.76$, $p = 0.017$). As we hypothesized, this effect was nonetheless restricted: There was a significant interaction between the group and the type of trustee ($F = 3.29$, $p = 0.038$). Participants who inhaled OT, as opposed to a placebo, transferred more money to reliable trustees, but did not transfer more money to unreliable trustees, revealing that OT does not increase trust when the partner appears unreliable (see Table 1 and Figure 2). The effect of OT on trust does not seem to be explainable by mood differences, subjective confidence, affection for human nature, or a perceived difference in condition assignment (all these variables were equal across conditions; see Supplementary Information).

This study is the first to demonstrate the boundary conditions of OT effects on pro-social behaviors. While our results are consistent with previous studies showing

that OT is a pro-social hormone[23], they nevertheless indicate that its effects are mitigated when pro-sociality is no longer adaptive. When hints exist that the partner might not be reliable, OT does not enhance trust.

Which mechanisms could account for OT's contextual effects? As already demonstrated by others[5,6], OT does not seem to alter the perception of the risk inherent to a situation. If this was the case, participants would have also transferred more money to unreliable partners after being administered OT. A much more likely explanation is that OT's effects are *moderated* by the perceived level of risk inherent to the interaction. Namely, OT effects would be maximal when the condition appears neutral or favorable (i.e., conditions in which an increase in trust is likely to bring about benefits), and nonexistent when the condition appears shady (i.e., conditions in which an increase in trust is could be detrimental). Thus, the higher the perceived risk, the lower the trust-enhancing effect of OT.

This hypothesis can be further tested using data from two previous studies[5,6], which found that oxytocin increased trust in humans but not in computers. If our hypothesis is valid, we should find evidence that participants in these studies perceived the computer condition as more risky than the human condition. Indeed, perhaps due to subtle differences in instruction (Heinrichs, personal communication), participants in *previous studies* considered it *more* risky to invest in a computer than in a human being (i.e., regardless of condition, mean transfers to the computer were lower than mean transfers to a human being; cf Table 1 in Kosfeld et al.'s main text[5], and Table 1 in Baumgartner et al.'s supplementary information[6]). Conversely, in *our study*, participants considered it *less* risky to invest in the computer than in a human being (i.e., regardless of condition, mean transfers to the computer were higher than mean transfers to a human being; see Table 1). Taken together, these results allow for much more specific predictions about when oxytocin will increase trust and when it

will not. Consistent with the moderating hypothesis stated above, it appears that oxytocin does not increase trust if conditions are deemed risky.

Our results have a number of important implications. First, they suggest that OT does not boost pro-social behaviour under all circumstances. Like most of our biological underpinnings, OT has been fine-tuned through natural selection to facilitate survival and adaptation. The fact that OT does so by enhancing peer bonding and interdependence does not preclude that it might have a different effect under conditions in which interdependence could prove harmful. Perhaps OT's effect could even reverse in particularly dangerous conditions. Second, this research shows that oxytocin is far from being the magical trust elixir described in the news, on the Internet, or even by some influential researchers[9]. Marketers, politicians, merchants, and others tempted to use oxytocin should be aware that it does not make people gullible. Our data suggest that OT will increase trust behaviors (e.g. investments, purchases, concessions made during negotiations) if the partner or the deal is perceived as neutral or trustworthy, but that it will not do so if the partner or if the deal looks suspicious.

Methods

Sixty healthy young adult men (M = 21.2, SD = 2.4) were enrolled in the study and randomly assigned to receive either intranasal placebo (PL; n = 30) or oxytocin (OT; n = 30; 32 IU Syntocinon Spray, Novartis, Basel, Switzerland). In order to avoid gender differences in OT response, only males were recruited for the study. Participants were informed at the time of enrolment that the experiment sought to investigate the effect of a hormone on cognitive and emotional processes. Before substance administration, participants filled in measures of demographics, risk taking, self-esteem, kindness, agreeableness, sociability, emotional competencies, and psychological symptoms, in order to ensure that groups were equal regarding all demographics and personality factors potentially relevant to the study.

The substance (OT or PL) was then inhaled. Owing to the role of social thoughts or experiences in triggering the effects of oxytocin, subjects were then invited to wait for the start of the experiment in front of an excerpt of a movie featuring friendship and camaraderie.

Forty-five minutes after substance inhalation, participants received written instructions for the trust game[21,22] (see Supplementary Information), which explained the rules of the game and the payment procedure at the end of the experiment. In one part of the game, participants were led to believe that they would play online with real people. Accordingly, they were provided with a brief description of their partner before each round (to ensure plausibility, subjects were also asked to provide such descriptions of themselves upon arrival at the laboratory). In fact, these descriptions were manipulated and pretested to induce either high or low trust (see Supplementary Information). Participants played each round of the trust game (depicted in Figure 1) with one of 10 different partners, of which 5 appeared trustworthy and 5 relatively untrustworthy. In another part of the game, participants were told that they would play 10 rounds with the computer, which would randomly determine the back-transfers. Participants did not receive any feedback about the back-transfers during the experiment. Before their leaving the laboratory, participants were asked to report on their beliefs about condition assignments, mood, trust and affection for human nature in order to control for confouding factors potentially associated with OT administration.

One outlier was removed, leaving 59 subjects for the analyses (29 in the oxytocin group and 30 in the placebo group). A 2 (condition: oxytocin or placebo) x 3 (Type of target: computer, human high trust, human low trust) mixed model was then performed on investments, with subject being a random factor, substance administered being a between-subject factor and truthworthiness of partner being a within-subject factor. Kindness, self-esteem, social competence, emotional competence and mental health were included as a covariates as they were found to

have an independent influence on investments. Significant ($p < .05$) multivariate effects were followed up with post-hoc tests with adjustment for multiple comparisons (Bonferroni).

[1] Ferguson, J.N. *et al.*, Social amnesia in mice lacking the oxytocin gene. *Nat. Genet.* **25**, 284-288 (2000).

[2] Ferguson, J.N., Young, L.J., & Insel, T.R., The neuroendocrine basis of social recognition. *Front. Neuroendocrinol.* **23**, 200-224 (2002).

[3] Winslow, J.T. & Insel, T.R., The social deficits of the oxytocin knockout mouse. *Neuropeptides* **36**, 221-229 (2002).

[4] Pedersen, C.A., Ascher, J.A., Monroe, Y.L., & Prange, A.J., Oxytocin induces maternal behavior in virgin female rats. *Science* **216**, 648-650 (1982).

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

[6] Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E., Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* **58**, 639-650 (2008).

[7] Theoridou, A., Rowe, A.C., Penton-Voak, I.S., & Rogers, P.J., Oxytocin and social perception: Oxytocin increases perceived facial trustworthiness and attractiveness. *Hormones and Behavior* **56**, 128-132 (2009).

[8] Zak, P.J., Stanton, A.A., & Ahmadi, S., Oxytocin increases generosity in humans. *PLoS One* **2** (2007).

[9] Damasio, A., Brain trust. *Nature* **435**, 571 (2005).

[10] Dethlefs, D.R. Chemically enhanced trust: Potential law enforcement and military applications for oxytocin. Naval Postgraduate School, Montery, CA (2007).

[11] Unkelbach, C., Guastella, A.J., & Forgas, J.P., Oxytocin selectively facilitates recognition of positive sex and relationship words. *Psycholol. Sci.* **19**, 1092-1094 (2008).

[12] Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S.C., Oxytocin improves “mind-reading” in humans. *Biol. Psychiatry* **61**, 731-733 (2007).

[13] Guastella, A.J., Mitchell, P.B., & Mathews, F., Oxytocin enhances the encoding of positive social memories in humans. *Biol. Psychiatry* **64**, 256-258 (2008).

[14] Taylor, S.E., Tend and befriend: Biobehavioral bases of affiliation under stress. *Current Directions in Psychological Science* **15**, 273-277 (2006).

[15] Hart, M., Buy oxytocin, Available at http://oxytocinreviews.us/Oxytocin/Buy-Oxytocin.html, (2008).

[16] Campbell, A., Attachment, aggression and affiliation: The role of oxytocin in female social behavior. *Biol. Psychol.* **77**, 1-10 (2008).

[17] Pedersen, C.A., Biological aspects of social bonding and the roots of human violence. *Ann. N. Y. Acad. Sci.* **1036**, 106–127 (2004).

[18] Debiec, J., Peptides of love and fear: vasopressin and oxytocin modulate the integration of information in the amygdala. *BioEssays* **27**, 869–873 (2005).

[19] Teoh, S.H. & Wong, T.J., Why new issues and high-accrual firms underperform: The role of analysts' credulity. *Review of Financial Studies* **15**, 869-900 (2002).

[20] Greenspan, S., Loughlin, G., & Black, R.S., Credulity and gullibility in persons with mental retardation: A framework for future Research. *International review of research in mental retardation* **24**, 101-135 (2001).

[21] Berg, J., Dickhaut, J., & McCabe, K., Trust, reciprocity, and social history. *Games and Economic Behavior* **10**, 122-142 (1995).

[22] Cesarini, D. *et al.*, Heritability of cooperative behavior in the trust game. *PNAS* **105** (10), 3721 (2008).

[23] Insel, T.R. & Fernald, R.D., How the brain processes social information: Searching for the Social Brain. *Annual Reviews of Neuroscience* **27**, 697-722 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgments This research was supported by the Belgian National Fund for Scientific Research. We thank Nathalie Lefèvre for statistical consultance and cécile Husquet and Noah Forrin for their proofreading work. We also thank Markus Heinrichs for sharing information that facilitated data interpretation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for material should be addressed to M.M. (moira.mikolajczak@uclouvain.be).

Author Contributions M.M. designed the experiment, analyzed the data, and prepared the manuscript. A.L. performed the experiment. P. d.T. took the medical responsibility for the study. O.L. and J.G. supervised the project at all stages. All authors commented on the manuscript.

Table 1 | Average Transfers as a Function of Group and Reliability of the Target

Type of Trustee	Group	Mean transfer (in EUR)	Standard error	Lower Bound (95% CI)	Upper Bound (95% CI)
Computer	Placebo	25.563	0.828	23.939	27.186
	Oxytocin	30.418	0.842	28.766	32.070
	All	27.990	0.588	23.959	27.607
Reliable partner	Placebo	24.326	1.305	21.764	26.887
	Oxytocin	27.240	1.328	24.635	29.846
	All	25.783	0.930	23.959	27.607
Unreliable partner	Placebo	15.876	1.305	13.314	18.437
	Oxytocin	15.103	1.328	12.497	17.708
	All	15.489	0.930	13.665	17.313

Note. Kindness, self-esteem, social competence, emotional competence and mental health were included as a covariates as they were found to have an independent influence on investments. Results hold with and without the inclusion of covariates, but the comparison of Akaike's Information Criteria (AIC) indicated that the fit of the model was better when covariates were included.

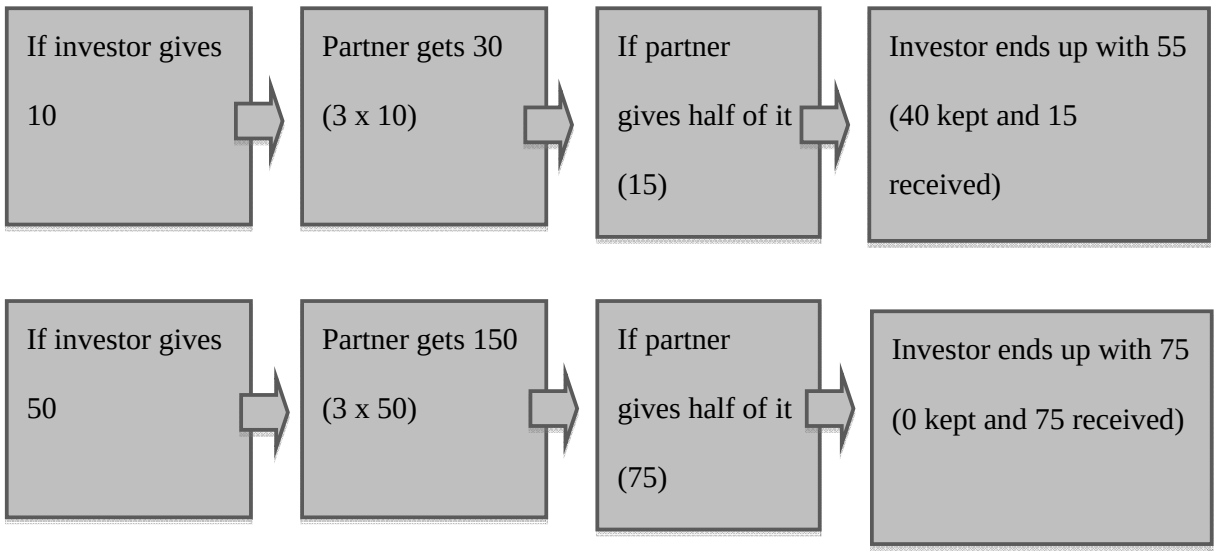

Figure 1 | The trust game. At the beginning of each round, the participant (i.e., investor) received an initial endowment of 50 EUR, of which he can send 0, 10, 20, 30, 40 or 50 to his partner. The experimenter tripled the sum transferred (which amounts to 0, 30, 60, 90, 120 or 150). The investor was informed that his partner can reimburse him any amount he wants, or nothing. For example, if the investor sent 30 EUR, the partner received 90 EUR and could choose any back transfer from 0 to 90 EUR. The back transfer was not tripled. Thus, the investor's final pay off amounted to the sum he did not transfer plus the back transfer from the partner. Each participant made 20 decisions: 5 with seemingly reliable partners, 5 with seemingly suspicious partners and 10 facing the computer.

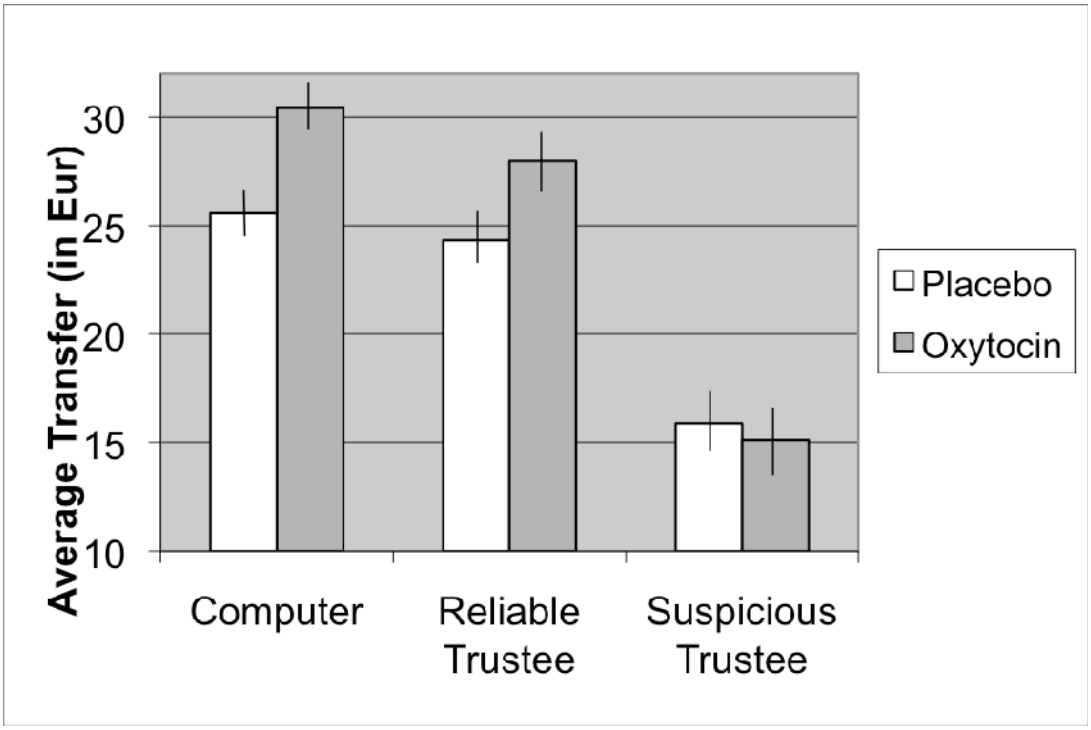

Figure 2 | Transfers as a Function of Group and Reliability of the Partner. Bars represent the average transfer (in EUR) that subjects made with each type of partner, in the placebo (white bars) and oxytocin (grey bars) conditions. Error bars represent the standard error of the mean. This Figure shows that oxytocin increases trust, except when the partner appears suspicious.

Supplementary Information

SI Tables

Supplementary Table 1: Means and standard deviations of demographic variables and individual differences measured before substance administration.

Variables	Placebo Means (and SDs)	Oxytocin Means (and SDs)	Independent samples t-test (and p-value)
Age	21.47 (2.73)	20.93 (2.07)	-0.85 (.40)
Body Mass Index	24.00 (3.13)	23.43 (5.21)	-0.52 (.61)
Risk Taking	2,51 (0.45)	2.65 (0.47)	1.14 (.26)
Kindness	2.98 (0.22)	2.98 (0.37)	0.00 (1.00)
Self-esteem	3.11 (0.44)	3.21 (0.55)	0.79 (.43)
Agreeableness	2.84 (0.39)	2.82 (0.41)	-0.13 (.90)
Social Competence	4.69 (0.47)	4.68 (0.53)	-0.15 (.89)
Emotional Competence	4.84 (0.60)	4.85 (0.58)	0.02 (.98)
Mental disorders	1.56 (0.51)	1.55 (0.47)	-0.07 (.94)

Note. These results show that groups were statistically equivalent regarding all the demographics and individual differences relevant to the study. BMI was computed as weight in kilograms/(height in meters)^2. Risk-taking, kindness, self-esteem, agreeableness, social competence, emotional competence and mental disorders were measured using the risk-

taking subscale of the Jackson Personality Inventory[24], the kindness subscale of the Value in Action scales[25], the Rosenberg Self-Esteem Scale[26], the agreeableness dimension of the NEO-PI_R[27], the trait Social Intelligence Questionnaire, the Trait Emotional Intelligence Questionnaire[28], and the Brief Symptom Inventory[29].

Supplementary Table 2: Means and standard deviations of variables measured after substance administration.

Variables	Placebo Means (and SDs)	Oxytocin Means (and SDs)	Independent samples t-test
Positive Mood	2.89 (0.45)	2.87 (0.48)	-0.22 (.83)
Trust in people	2.53 (0.43)	2.43 (0.37)	-0.89 (.38)
Affection for human nature	2.62 (0.53)	2.51 (0.59)	-0.75 (.46)

Note. These results suggest that the effect of oxytocin on trust behavior is not due to differences in mood, self-reported trust, or affection for human nature. Positive mood, trust and affection for the human nature were measured using questionnaires developed for the purpose of the present study.

SI Methods

Instructions for the trust game (adapted from Cesarini et al. 2008[23])

"*In this section you will be randomly paired with several other participants (a different one at each round). You will not find out who these people are, nor will they find out who you are, not now, nor after the experiment is over. The only information you will be given before each round is the first name of your partner, his age, his faculty and his main hobby. Please raise your hand if you have the feeling that you might know that person*" [In order to ensure maximum plausibility, each participant was required to give his age, faculty and main hobby to the experimenter at the beginning of the study].

"*You will receive 50 EUR at the beginning of each round, and your task is to decide what share of these 50 EUR to transfer to the person in the other room. The money you give to your partner is tripled; in other words, for every 10 EUR you decide to transfer, your partner receives 30. Your partner will then decide how much of the (tripled) money to return to you. He can send any amount between zero and his total amount available back to you.You will then earn whatever money is returned to you plus the share of the 50 EUR you decided to keep. To assist you in your decision, the table next to you shows how much money your partner receives depending on how much you decide to transfer.*

Amount you decide to transfer (in EUR)	Amount your partner receives (in EUR)
0	0
10	30

20	60
30	90
40	120
50	150

"*You will not know the amount your partners have decided to transfer back to you until the end of the experiment. Please keep in mind that what you earned in this game is really what you might earn from your participation in this study. As we told you, one participant will indeed be randomly selected at the end of the experiment to receive the total amount of his back transfers in cash*" [Accordingly, we rewarded the best performer with 300 EUR at the end of the study]."

Participants were then provided with an oral summary of the rules of the game, which were illustrated by several examples (such as those represented in Figure 1). All subjects understood the explanations.

Pretest of the trustworthiness of the targets

Partners' descriptions for the trust game were manipulated to induce trust or mistrust. As each description only contained the partner's first name, age, education and main hobby, trust or mistrust was induced on these characteristics only. Trust level inspired by different educations and hobbies was pretested on 20 participants. A pretest highlighted that participants would trust psychology or philosophy students more than marketing or political science students. Similarly, hobbies such as youth movements or first-aid were more trustworthy than hobbies like gambling or violent sports. This trust manipulation appears to have been effective, as we observed a main effect of the type of partner ($p < 0.001$). Average investment in a trust-inspiring partner was 25.78 EUR while it was down to 15.49 EUR in an unreliable partner.

SI References

[24] Jackson, D. N. Jackson Personality Inventory-Revised (Manual) (Sigma Assessment Systems, Port Huron, MI, 1994).

[25] Park, N., Peterson, C. & Seligman, M. E. P. Strengths of character and well-being. J. Soc. Clin. Psychol. **23**, 603-619 (2004).

[26] Rosenberg, M. Rosenberg self-esteem scale (Basic Books, New York, 1979).

[27] Costa, P. T. & McCrae, R. R. Neo PI-R Professional Manual (Psychological Assessment Resources, Odessa, FL, 1992).

[28] Petrides, K. V. & Furnham, A. Trait emotional intelligence: Behavioural validation in two studies of emotion recognition and reactivity to mood induction. Eur. J. Personality **17**, 39-57 (2003).

[29] Derogatis, L. R. BSI: Brief Symptom Inventory (Manual) (National Computer Systems, Minneapolis, 1993).