

Distinct oxytocin effects on belief updating in response to desirable and undesirable feedback

Yina Ma^{a,1}, Shiyi Li^a, Chenbo Wang^b, Yi Liu^b, Wenxin Li^b, Xinyuan Yan^a, Qiang Chen^c, and Shihui Han^{b,1}

^aState Key Laboratory of Cognitive Neuroscience and Learning, International Data Group (IDG)/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China; ^bSchool of Psychological and Cognitive Sciences, IDG/McGovern Institute for Brain Research, Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100080, China; and ^cLieber Institute for Brain Development, Baltimore, MD 21205

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved June 27, 2016 (received for review March 17, 2016)

Humans update their beliefs upon feedback and, accordingly, modify their behaviors to adapt to the complex, changing social environment. However, people tend to incorporate desirable (better than expected) feedback into their beliefs but to discount undesirable (worse than expected) feedback. Such optimistic updating has evolved as an advantageous mechanism for social adaptation. Here, we examine the role of oxytocin (OT)—an evolutionary ancient neuropeptide pivotal for social adaptation—in belief updating upon desirable and undesirable feedback in three studies ($n = 320$). Using a double-blind, placebo-controlled between-subjects design, we show that intranasally administered OT (IN-OT) augments optimistic belief updating by facilitating updates of desirable feedback but impairing updates of undesirable feedback. The IN-OT-induced impairment in belief updating upon undesirable feedback is more salient in individuals with high, rather than with low, depression or anxiety traits. IN-OT selectively enhances learning rate (the strength of association between estimation error and subsequent update) of desirable feedback. IN-OT also increases participants' confidence in their estimates after receiving desirable but not undesirable feedback, and the OT effect on confidence updating upon desirable feedback mediates the effect of IN-OT on optimistic belief updating. Our findings reveal distinct functional roles of OT in updating the first-order estimation and second-order confidence judgment in response to desirable and undesirable feedback, suggesting a molecular substrate for optimistic belief updating.

oxytocin | social adaptation | confidence | belief updating | optimism

Humans learn from their experiences and adaptively update their beliefs and behaviors in response to the complex, changing social environment. This process of belief updating is essential for social adaptation and survival. However, people tend to incorporate desirable (better than expected) feedback into their beliefs but to discount undesirable (worse than expected) feedback. Such optimistic updating has evolved as an advantageous mechanism for social adaptation. Here, we examine the role of oxytocin (OT)—an evolutionary ancient neuropeptide pivotal for social adaptation—in belief updating upon desirable and undesirable feedback in three studies ($n = 320$). Using a double-blind, placebo-controlled between-subjects design, we show that intranasally administered OT (IN-OT) augments optimistic belief updating by facilitating updates of desirable feedback but impairing updates of undesirable feedback. The IN-OT-induced impairment in belief updating upon undesirable feedback is more salient in individuals with high, rather than with low, depression or anxiety traits. IN-OT selectively enhances learning rate (the strength of association between estimation error and subsequent update) of desirable feedback. IN-OT also increases participants' confidence in their estimates after receiving desirable but not undesirable feedback, and the OT effect on confidence updating upon desirable feedback mediates the effect of IN-OT on optimistic belief updating. Our findings reveal distinct functional roles of OT in updating the first-order estimation and second-order confidence judgment in response to desirable and undesirable feedback, suggesting a molecular substrate for optimistic belief updating.

Humans learn from their experiences and adaptively update their beliefs and behaviors in response to the complex, changing social environment. This process of belief updating is essential for social adaptation and survival. However, people tend to incorporate desirable (better than expected) feedback into their beliefs but to discount undesirable (worse than expected) feedback. Such optimistic updating has evolved as an advantageous mechanism for social adaptation. Here, we examine the role of oxytocin (OT)—an evolutionary ancient neuropeptide pivotal for social adaptation—in belief updating upon desirable and undesirable feedback in three studies ($n = 320$). Using a double-blind, placebo-controlled between-subjects design, we show that intranasally administered OT (IN-OT) augments optimistic belief updating by facilitating updates of desirable feedback but impairing updates of undesirable feedback. The IN-OT-induced impairment in belief updating upon undesirable feedback is more salient in individuals with high, rather than with low, depression or anxiety traits. IN-OT selectively enhances learning rate (the strength of association between estimation error and subsequent update) of desirable feedback. IN-OT also increases participants' confidence in their estimates after receiving desirable but not undesirable feedback, and the OT effect on confidence updating upon desirable feedback mediates the effect of IN-OT on optimistic belief updating. Our findings reveal distinct functional roles of OT in updating the first-order estimation and second-order confidence judgment in response to desirable and undesirable feedback, suggesting a molecular substrate for optimistic belief updating.

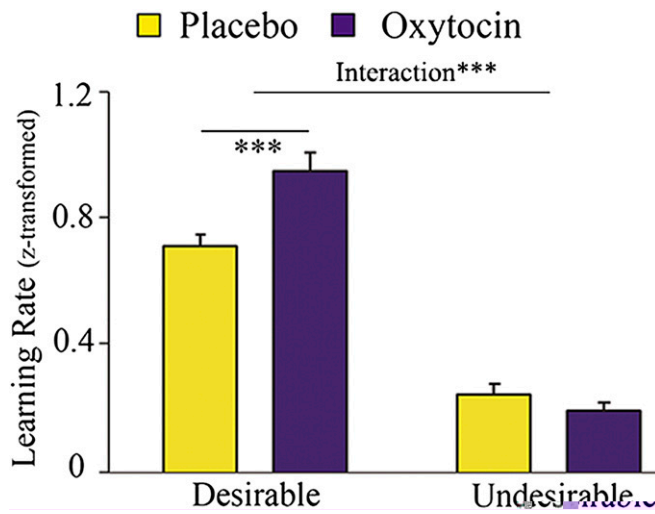


Fig. 3. IN-OT enhanced learning rate related to desirable but not undesirable feedback. *** $P < 0.001$.

...min. h c. l. in h i. W. n h. i. m. i. i. in
 ...in (B, in B, D - B, U, n) n. m. c.
 ...n. / ... n. h. B. W. n. i. n. i. n. l.
 ...l. n. l. w. i. h. C. U. D. ($r = 0.328, P < 0.001$). Th. c. W. n.
 ...c. l. l. r. l. i. n. W. n. B. n. C. U. D. ($r =$
 $0.110, P = 0.245$) c. W. n. B. n. n. v. r. l.
 $(r = -0.026, P = 0.780)$ c. n. r. l. ($r = -0.013, P =$
 0.887). h. n. n. m. h. m. i. n. l. i. (SI Appendix, SI
 Methods) ... m. h. h. T. m. n. i. n. T. m. n. W.
 ...n. n. i. n. n. i. n. B. W. m. i. i.
 ...: $t = 2.36, P = 0.018$; SI Appendix, Fi. 4C n. T. l. 12-
 15). Th. W. i. c. i. n. l. i. n. T. m. n. W.
 l. n. c. i. n. i. n. W. h. n. i. n. h. c. W. i. h. C. U. D., $B = 2.76,$
 $t(111) = 1.42, P = 0.157,$ m. c. W. i. h. i. n. i. l. v. i. n., $B =$
 $4.67, t(112) = 2.46, P = 0.016,$ i. n. h. h. T. n.
 C. U. D. ... l. l. m. i. c. h. T. n. B. A.
 ... m. i. n. l. i. (SI Appendix, SI Methods) h.
 ... i. n. n. l. i. m. i. n. W. n. c. m.
 ... r. W. i. h. 95% n. i. n. (n. i. n. i. n. c. l.: 0.58-4.32).

Matched Mood and Trait Between OT and PL Groups. T. n. L.
 ... n. n. i. n. c. i. i. m. m. n. i. ,
 ... c. i. n. c. i. n. c. i. n. c. m. m. l. c.
 ... n. h. c. i. i. (SI Appendix, T. l. 2 n. 16-18).
 M. c. c. n. i. h. c. i. n. m. m. c. c. m. n. n. c. i. n.
 i. m. c. i. n. r. n. n. i. m. i. n. c. i. n. i. n. l.
 W. n. T. n. L. c. (SI Appendix, T. l. 19 n. 20),
 ... i. n. h. h. I. T. n. i. n. i. n. n.
 ... c. i. n. T. i. n. h. h. n. i. n. i. i. (i. e., c.
 ... i. n. i. m. , m. m. c. c. m. n. n. c. i. n.) .

Discussion

Th. c. i. n. c. i. l. n. n. n. i. n. h. i. c.
 ... c. i. n. l. c. i. l. n. i. n. i. n. h. n. i. n.
 ... n. i. m. n. i. m. i. i. n. h. l. n. i. n. n. i.
 ... m. h. n. i. m. c. h. i. l. n. m. n. l. h. l. h. (1, 2, 26-28). H. c. W.
 ... h. W. i. n. c. i. n. n. i. m. T. n. i. m. i. i.
 ... i. n. ... i. l. l. W. m. n. c. h. I. T. i. n. c.
 ... i. n. i. n. c. n. n. i. l. c. i. n. c. i. n. c.
 ... i. n. n. n. r. l. c. i. n. Th. i. n. T. n. n.
 ... i. n. W. c. l. i. n. n. h. l. c. i. n. c. , i. e., T.
 ... l. i. l. i. f. c. i. n. l. c. i. n. c. m. i. l.
 ... n. n. r. l. c. i. n. c. c. h. r. l. c.
 ... i. n. i. n. m. l. m. n. c. i. i. n. n. T. n. h.

... i. l. i. n. l. (10-15) n. c. i. n. h. T. i. m.
 ... n. n. i. n. i. c. i. n. c. i. n. c. i. n.
 ... n. i. n. c. c. l. c. c. h. T. i. m. l. l. c.
 ... c. i. n. c. i. m. i. i. i. n. n. l. i. n.
 ... n. i. n. l. l. i. n. i. n. n. n. r. l. c.
 ... n. r. l. c. i. n. h. I. T. (L) i. n. i. n. l. n.
 ... i. m. i. n. i. m. n. m. m. c. i. n. h. h. T.
 ... n. i. m. i. i. n. W. c. n. T. i. n. n. c. l. T.
 ... n. n. i. n. c. i. n. i. i. l. Th. c. l. W. c. i. n. l.
 ... W. i. h. c. i. i. n. i. n. h. i. m. i. i. n. n. c. i. n.
 ... i. n. c. c. l. n. h. i. l. i. n. i. n. , n. i. i. ,
 ... m. m. n. i. i. i. n. c. i. n. r. l. n. n. r. l. c.
 ... (19, 20, 45), c. i. c. n. l. c. i. n. c. i. n. l. n.
 ... m. c. i. n. c. i. n. i. n. c. i. n. (20, 41). I. h. n. c. i. n.
 ... h. h. n. c. i. n. i. n. c. i. n. W. i. c. c. l. i. h. n. W.
 ... c. i. n. h. W. i. c. i. c. i. c. c. m. (47). Th. m. c.
 ... m. i. n. n. i. n. c. c. i. n. i. n. c. i. n. i. n. h. c. n. c.
 ... h. i. m. i. i. h. i. n. c. c. (41). C. n. i. n. W. i. h. h. i.
 ... c. i. i. n. h. W. h. h. T. n. i. m. i. i. n. i. n.
 ... W. m. i. h. h. T. n. n. n. i. n. i. n. n.
 ... r. l. c. i. n. n. i. l. m. h. n. i. m. n. c. l. i. n.
 ... T. i. i. i. m. i. i. n. l. T. m. i. h. i. n. c. i. n. i.
 ... i. l. c. i. n. c. i. n. h. c. (i. e., n. c. c. n),
 ... h. i. n. h. r. i. n. h. n. i. m. i. n. W. i. h. m. c.
 ... n. i. n. i. l. l. i. n. h. r. l. n. i. n.
 ... Th. i. n. T. h. h. i. m. h. n. i. m.
 ... n. c. l. i. n. T. n. i. l. n. i. n. (5, 33) h. h. W. c.
 ... W. i. c. c. c. n. T. n. i. n. i. n. r. l. n.
 ... n. r. l. n. F. r. m. l. h. i. l. m. i. n. h. h.
 ... i. W. i. h. c. h. T. m. i. n. c. i. n. i. c. W. c. m.
 ... i. l. i. n. c. i. n. (48), c. h. I. T. W. i. i. i.
 ... i. n. n. r. l. c. i. n. i. l. n. n. i. n.
 ... n. n. r. l. c. i. n. Th. i. l. i. n. h. h. i. W. i. h.
 ... n. h. T. n. h. n. i. n. i. n. i. l. l.
 ... i. n. n. n. l. l. n. (33, 49), c. h. T. W. i. n. c.
 ... l. c. i. n. n. i. n. h. r. l. n. n. r. l. c.
 ... Th. r. n. i. n. i. n. T. n. i. n. i. n.
 ... n. r. l. c. n. r. l. c. n. n. l. i. n.
 ... h. h. h. Th. i. l. i. n. m. l. (5), W. i. h. c.
 ... h. h. c. i. n. i. c. i. n. n. h. n. i. n. i. i.
 ... c. i. n. c. i. i. l. i. n. n. c. m. i. l. l.
 ... i. n. c. h. I. T. i. l. c. i. n. n. i. n.
 ... r. l. c. i. n. i. m. i. n. h. l. c. i. n. n. i. n.
 ... r. l. c. i. n. Th. c. i. n. h. T. i. n. c. i. n.
 ... c. m. i. l. c. i. n. c. m. n. r. l.
 ... i. W. i. h. h. i. l. i. n. m. l. (5).
 ... A. m. l. i. n. i. n. h. h. W. n. c. n. c. T. i. n. i.
 ... i. l. W. i. h. h. c. i. n. i. (34), h. h. i. c. i. (9), i. m.
 ... i. m. i. n. c. l. i. n. (35), l. W. m. i. n. l. n. i. i. (36),
 ... h. h. h. m. n. i. n. (50). I. T. l. n. i. n. i. n. i. n. l.
 ... W. i. h. m. l. n. c. c. c. h. n. i. c. c. (51), i. m. (52),
 ... n. n. c. i. n. (24, 25). Th. i. l. i. n. m. l. T. n. i. n.
 ... c. i. n. c. c. T. n. i. l. c. i. n. l. l. i. l.
 ... c. i. n. i. l. (5). I. n. c. h. i. m. l. W. h. h.
 ... T. r. n. i. n. n. n. r. l. c. i. n. i. l.
 ... i. l. c. i. n. l. l. i. l. i. n. i. l. (i. e., h.
 ... W. i. h. h. r. c. i. n. c. i. n. i. c. i). Th. i. n. h. I. T.
 ... n. c. m. i. h. h. c. c. n. n. r. l. c. i. n.
 ... l. c. i. n. i. n. i. l. h. i. m. c. n. i. l. i. m. i. n. c.
 ... T. r. m. i. n. c. i. n. W. i. h. h. i. c. i. h. l.
 ... i. m. i. i. n. (2, 20). A. l. h. I. T. Th. c. n.
 ... i. i. c. i. n. i. n. W. i. n. i. l. c. i. l. (24, 25), h. n. i.
 ... m. h. n. i. m. n. c. l. i. n. h. n. i. l. h. c. i. i. T.
 ... i. n. c. c. i. n. c. m. i. n. l. c. c. i. n. i. n.
 ... n. i. l. n. i. m. h. n. i. m. h. W. i. h. I. T. m. i. i.
 ... c. i. n. i. m. i. n. c. i. n.
 ... H. W. c. , m. c. i. h. h. W. n. c. n. c. T. -
 ... i. n. m. c. i. l. l. i. n. i. l. h. h. W. i. h. i. W.
 ... h. m. n. i. (53) c. l. W. i. l. n. i. (54). I. h. n.
 ... h. c. i. l. c. c. i. n. i. n. h. i. n. i. n.

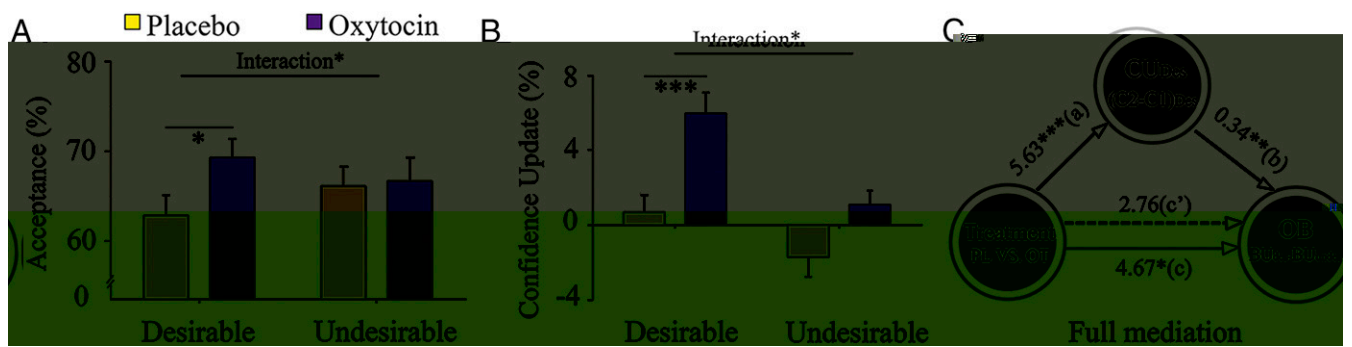


Fig. 4. (A) IN-OT increased participants' acceptance of desirable (but not undesirable) feedback. (B) OT increased participants' confidence in their estimates after receiving desirable but not undesirable feedback. (C) Moreover, the OT effect on optimistic bias (OB) in belief updating was mediated by the effect of OT on confidence update upon desirable feedback (CU_{des}). **P* < 0.05; ***P* < 0.01; ****P* < 0.001.

... (54). The ... (53). ... (55). ... (56). ... (57). ... (58). ... (59). ... (60). ... (61). ... (62). ... (63). ... (64). ... (65). ... (66). ... (67). ... (68). ... (69). ... (70). ... (71). ... (72). ... (73). ... (74). ... (75). ... (76). ... (77). ... (78). ... (79). ... (80). ... (81). ... (82). ... (83). ... (84). ... (85). ... (86). ... (87). ... (88). ... (89). ... (90). ... (91). ... (92). ... (93). ... (94). ... (95). ... (96). ... (97). ... (98). ... (99). ... (100).

... (101). ... (102). ... (103). ... (104). ... (105). ... (106). ... (107). ... (108). ... (109). ... (110). ... (111). ... (112). ... (113). ... (114). ... (115). ... (116). ... (117). ... (118). ... (119). ... (120). ... (121). ... (122). ... (123). ... (124). ... (125). ... (126). ... (127). ... (128). ... (129). ... (130). ... (131). ... (132). ... (133). ... (134). ... (135). ... (136). ... (137). ... (138). ... (139). ... (140). ... (141). ... (142). ... (143). ... (144). ... (145). ... (146). ... (147). ... (148). ... (149). ... (150).

Methods

Ethics Approval. The experimental procedures were in line with the standards set by the Declaration of Helsinki and were approved by the local Research Ethics Committee of the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University. Participants provided written informed consent after the experimental procedure had been fully explained and were reminded of their right to withdraw at any time during the study.

Participants. We recruited 320 male Chinese college students as paid volunteers. Twelve participants (3.75%) were dropped from data analysis because of technical problems or participants' failure to complete the study. Data from 308 participants were included in the final data analysis: 99 participants in study 1 (50 under PL, 49 under OT), 95 participants in study 2 (47 under PL, 48 under OT), and 114 participants in study 3 (57 under PL, 57 under OT). All participants reported no history of neurological or psychiatric diagnoses. Exclusion criteria were self-reported medical or psychiatric disorder and drug/alcohol abuse. Participants were instructed to refrain from smoking or drinking (except water) for 2 h before the experiment.

Procedure. All three studies were conducted by following a randomized, placebo-controlled, double-blind, between-subjects design. Participants first completed a set of questionnaires and were then administered with OT or PL and performed the belief updating task 40 min later. The procedure of OT and PL administration was similar to previous work (15–17). A single intranasal dose of 24 IU OT or PL (containing the same ingredients

except for the neuropeptide) was self-administered by nasal spray under experimenter supervision. Finally, participants completed the mood measurement again.

The Belief Update Task. In studies 1 and 2, participants completed two sessions of life event estimation. Participants were first presented with 40 different adverse life events (*SI Appendix, SI Methods*) and estimated their likelihood (0–99%) of experiencing each event on a self-paced basis (first Estimate). Participants were then presented with the probability of each event occurring to an average person in a similar environment (Feedback). Five minutes after the first session, participants were invited to complete a second estimation session, in which participants were presented with these 40 events in a random order and estimated the likelihood of each event again (second Estimate). The number of desirable and undesirable trials was reported in *SI Appendix, Table S21*. After the second session, participants were given a

surprise memory test for the presented feedback. The belief update task in study 3 was similar to that in studies 1 and 2, except that, for each event, participants additionally made judgment of (i) confidence in their first and second Estimate; and (ii) acceptance of the presented feedback.

ACKNOWLEDGMENTS. We acknowledge Drs. M. Crockett, A. Kappes, S. Shamay-Tsoory, and C. Zink for their valuable comments on an early draft; B. Li for OT preparation; and Dr. K. Woodcock for proofreading. This work was supported by startup funding from the State Key Laboratory of Cognitive Neuroscience and Learning, IDG/McGovern Institute for Brain Research; Open Research Fund of the State Key Laboratory of Cognitive Neuroscience and Learning (Y.M.); Beijing Municipal Science & Technology Commission (Z151100003915122) (to Y.M.); National Natural Science Foundation of China Projects 31470986, 31421003, and 91332125; and Ministry of Education of China Project 20130001110049 (to S.H.).

- McKay RT, Dennett DC (2009) The evolution of misbelief. *Behav Brain Sci* 32(6): 493–510, discussion 510–561.
- Sharot T (2011) The optimism bias. *Curr Biol* 21(23):R941–R945.
- Carter CS (2014) Oxytocin pathways and the evolution of human behavior. *Annu Rev Psychol* 65:17–39.
- Ishak WW, Kahloun M, Fakhry H (2011) Oxytocin role in enhancing well-being: A literature review. *J Affect Disord* 130(1–2):1–9.
- Ma Y, Shamay-Tsoory S, Han S, Zink CF (2016) Oxytocin and social adaptation: Insights from neuroimaging studies of healthy and clinical populations. *Trends Cogn Sci* 20(2): 133–145.
- Domes G, Heinrichs M, Michel A, Berger C, Herpertz SC (2007) Oxytocin improves “mind-reading” in humans. *Biol Psychiatry* 61(6):731–733.
- Riem MME, Bakermans-Kranenburg MJ, Voorthuis A, van IJendoorn MH (2014) Oxytocin effects on mind-reading are moderated by experiences of maternal love withdrawal: An fMRI study. *Prog Neuropsychopharmacol Biol Psychiatry* 51:105–112.
- Radke S, de Bruijn ERA (2015) Does oxytocin affect mind-reading? A replication study. *Psychoneuroendocrinology* 60:75–81.
- Bartz JA, et al. (2010) Oxytocin selectively improves empathic accuracy. *Psychol Sci* 21(10):1426–1428.
- Guastella AJ, Mitchell PB, Mathews F (2008) Oxytocin enhances the encoding of positive social memories in humans. *Biol Psychiatry* 64(3):256–258.
- Marsh AA, Yu HH, Pine DS, Blair RJ (2010) Oxytocin improves specific recognition of positive facial expressions. *Psychopharmacology (Berl)* 209(3):225–232.
- Guastella AJ, Mitchell PB, Dadds MR (2008) Oxytocin increases gaze to the eye region of human faces. *Biol Psychiatry* 63(1):3–5.
- Unkelbach C, Guastella AJ, Forgas JP (2008) Oxytocin selectively facilitates recognition of positive sex and relationship words. *Psychol Sci* 19(11):1092–1094.
- De Dreu CKW, et al. (2010) The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science* 328(5984):1408–1411.
- Ma Y, Liu Y, Rand DG, Heatherton TF, Han S (2015) Opposing oxytocin effects on intergroup cooperative behavior in intuitive and reflective minds. *Neuropsychopharmacology* 40(10):2379–2387.
- Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E (2005) Oxytocin increases trust in humans. *Nature* 435(7042):673–676.
- 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.
- Nave G, Camerer C, McCullough M (2015) Does oxytocin increase trust in humans? A critical review of research. *Perspect Psychol Sci* 10(6):772–789.
- Sharot T, Korn CW, Dolan RJ (2011) How unrealistic optimism is maintained in the face of reality. *Nat Neurosci* 14(11):1475–1479.
- Korn CW, Sharot T, Walter H, Heekeren HR, Dolan RJ (2014) Depression is related to an absence of optimistically biased belief updating about future life events. *Psychol Med* 44(3):579–592.
- Eil D, Rao JM (2011) The good news-bad news effect: Asymmetric processing of objective information about yourself. *Am Econ J Microecon* 3(2):114–138.
- Choe HK, et al. (2015) Oxytocin mediates entrainment of sensory stimuli to social cues of opposing valence. *Neuron* 87(1):152–163.
- Saphire-Bernstein S, Way BM, Kim HS, Sherman DK, Taylor SE (2011) Oxytocin receptor gene (OXTR) is related to psychological resources. *Proc Natl Acad Sci USA* 108(37):15118–15122.
- MacDonald K, et al. (2012) Oxytocin and psychotherapy: A pilot study of its physiological, behavioral and subjective effects in males with depression. *Psychoneuroendocrinology* 38(12):2831–2843.
- Mercedes Perez-Rodriguez M, Mahon K, Russo M, Ungar AK, Burdick KE (2015) Oxytocin and social cognition in affective and psychotic disorders. *Eur Neuropsychopharmacol* 25(2):265–282.
- Taylor SE, Kemeny ME, Reed GM, Bower JE, Gruenewald TL (2000) Psychological resources, positive illusions, and health. *Am Psychol* 55(1):99–109.
- Taylor SE, Broffman JI (2011) Psychosocial resources: Functions, origins, and links to mental and physical health. *Adv Exp Soc Psychol* 44:1–57.
- Carver CS, Scheier MF (2014) Dispositional optimism. *Trends Cogn Sci* 18(6):293–299.
- Vollmann M, Antoniw K, Hartung FM, Renner B (2011) Social support as mediator of the stress buffering effect of optimism: The importance of differentiating the recipients’ and providers’ perspective. *Eur J Pers* 25(2):146–154.
- Andersson MA (2012) Dispositional optimism and the emergence of social network diversity. *Sociol Q* 53(1):92–115.
- Ruiz JM, Matthews KA, Scheier MF, Schulz R (2006) Does who you marry matter for your health? Influence of patients’ and spouses’ personality on their partners’ psychological well-being following coronary artery bypass surgery. *J Pers Soc Psychol* 91(2):255–267.
- Taylor ZE, et al. (2012) Dispositional optimism: A psychological resource for Mexican-origin mothers experiencing economic stress. *J Fam Psychol* 26(1):133–139.
- Bartz JA, Zaki J, Bolger N, Ochsner KN (2011) Social effects of oxytocin in humans: Context and person matter. *Trends Cogn Sci* 15(7):301–309.
- Alvares GA, Chen NTM, Balleine BW, Hickie IB, Guastella AJ (2012) Oxytocin selectively moderates negative cognitive appraisals in high trait anxious males. *Psychoneuroendocrinology* 37(12):2022–2031.
- Quirin M, Kuhl J, Düsing R (2011) Oxytocin buffers cortisol responses to stress in individuals with impaired emotion regulation abilities. *Psychoneuroendocrinology* 36(6):898–904.
- Leknes S, et al. (2013) Oxytocin enhances pupil dilation and sensitivity to ‘hidden’ emotional expressions. *Soc Cogn Affect Neurosci* 8(7):741–749.
- Puskar KR, Sereika SM, Lamb J, Tusaie-Mumford K, McGuinness T (1999) Optimism and its relationship to depression, coping, anger, and life events in rural adolescents. *Issues Ment Health Nurs* 20(2):115–130.
- Hirsch JK, Walker KL, Chang EC, Lyness JM (2012) Illness burden and symptoms of anxiety in older adults: Optimism and pessimism as moderators. *Int Psychogeriatr* 24(10):1614–1621.
- Lebreton M, Abitbol R, Daunizeau J, Pessiglione M (2015) Automatic integration of confidence in the brain valuation signal. *Nat Neurosci* 18(8):1159–1167.
- De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice. *Nat Neurosci* 16(1):105–110.
- Sharot T, Garrett N (2016) Forming beliefs: Why valence matters. *Trends Cogn Sci* 20(1):25–33.
- Ma WJ, Jazayeri M (2014) Neural coding of uncertainty and probability. *Annu Rev Neurosci* 37:205–220.
- Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961) An inventory for measuring depression. *Arch Gen Psychiatry* 4(6):561–571.
- Spielberger CD, Gorsuch RL (1983) *State-Trait Anxiety Inventory for Adults: Manual, Instrument, and Scoring Guide* (Mind Garden, Inc., Menlo Park, CA).
- Garrett N, et al. (2014) Losing the rose tinted glasses: Neural substrates of unbiased belief updating in depression. *Front Hum Neurosci* 8:639.
- Palmiter S, et al. (2012) Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron* 76(5):998–1009.
- Vilares I, Howard JD, Fernandes HL, Gottfried JA, Kording KP (2012) Differential representations of prior and likelihood uncertainty in the human brain. *Curr Biol* 22(18):1641–1648.
- Stavropoulos KK, Carver LJ (2013) Research review: Social motivation and oxytocin in autism—implications for joint attention development and intervention. *J Child Psychol Psychiatry* 54(6):603–618.
- Shamay-Tsoory SG, Abu-Akel A (2016) The social salience hypothesis of oxytocin. *Biol Psychiatry* 79(3):194–202.
- De Dreu CK (2012) Oxytocin modulates the link between adult attachment and cooperation through reduced betrayal aversion. *Psychoneuroendocrinology* 37(7):871–880.
- Labuschagne I, et al. (2010) Oxytocin attenuates amygdala reactivity to fear in generalized social anxiety disorder. *Neuropsychopharmacology* 35(12):2403–2413.
- Watanabe T, et al. (2014) Mitigation of sociocommunicational deficits of autism through oxytocin-induced recovery of medial prefrontal activity: A randomized trial. *JAMA Psychiatry* 71(2):166–175.
- Bartz JA, et al. (2010) Effects of oxytocin on recollections of maternal care and closeness. *Proc Natl Acad Sci USA* 107(50):21371–21375.
- Radke S, Roelofs K, de Bruijn ERA (2013) Acting on anger: Social anxiety modulates approach-avoidance tendencies after oxytocin administration. *Psychol Sci* 24(8): 1573–1578.
- Varki A (2009) Human uniqueness and the denial of death. *Nature* 460(7256):684.
- Weinstein ND, Klein WM (1995) Resistance of personal risk perceptions to behaving interventions. *Health Psychol* 14(2):132–140.
- Lukas M, et al. (2011) The neuropeptide oxytocin facilitates pro-social behavior and prevents social avoidance in rats and mice. *Neuropsychopharmacology* 36(11):2159–2168.
- Toth I, Neumann ID, Slattery DA (2012) Central administration of oxytocin receptor ligands affects fear extinction in rats and mice in a timepoint-dependent manner. *Psychopharmacology (Berl)* 223(2):149–158.
- Sharot T, Guitart-Masip M, Korn CW, Chowdhury R, Dolan RJ (2012) How dopamine enhances an optimism bias in humans. *Curr Biol* 22(16):1477–1481.

Supporting Information

Distinct oxytocin effects on belief updating in response to desirable and undesirable feedback

Yina Ma¹, Shiyi Li¹, Chenbo Wang², Yi Liu², Wenxin Li², Xinyuan Yan¹, Qiang Chen³,

Shihui Han²

¹State Key Laboratory of Cognitive Neuroscience and Learning,
International Data Group (IDG)/McGovern Institute for Brain Research,
Beijing Normal University, Beijing, 100875, China

²School of Psychological and Cognitive Sciences, IDG/McGovern Institute for Brain
Research, Beijing Key Laboratory of Behavior and Mental Health, Peking University,
Beijing, 100080, China

³Lieber Institute for Brain Development, Baltimore, MD 21205, USA

Running title: Oxytocin and belief updating

Number of figures: 4

Supporting information: 21 Tables, and 6 Figures

Correspondence should be addressed to:

Yina Ma Ph. D.

State Key Laboratory of Cognitive Neuroscience and Learning,

Beijing Normal University,

19 Xin Jie Kou Wai Da Jie, Beijing, 100875, China

Phone/Fax: 8610-5880-2846

Email: yma@bnu.edu.cn

or

Shihui Han Ph. D.

School of Psychological and Cognitive Sciences

Peking University, 52 Haidian Street, Beijing 100080, China

Email: shan@pku.edu.cn

Supporting Methods

Pilot study to determine feedback for main experiments

The pilot study recruited 40 participants (15 males, mean age = 23.0 year, SD = 3.7). Participants were asked to estimate the probability (from 0 to 99%) of 100 different adverse life events that may happen to an average individual living in a similar socio-cultural environment. Eighty events were selected from the stimulus list of the previous study¹ and 20 additional events were complemented in the current study. Since all participants in the current study were college students, we asked participants to estimate the likelihood of these events occurring to an average Chinese college student. We also asked participants to identify those among the 100 life events that: 1) they had never heard of or did not understand; and 2) they were experiencing, or had experienced. An item was excluded if more than 5% of the participants had never heard of it, or did not understand it, or if more than 70% of the participants had experienced or were experiencing it. Forty-four adverse life events (e.g., “cancer”, “obesity”, “unemployed”, “depression”, “divorce” etc.) were randomly selected from the current stimulus set. Four adverse life events were used for practice and 40 adverse life events were used in the main experiments. The mean probability rating score of each event occurring to an average person obtained in this study was then used as social feedback in the main experiments.

Questionnaire measurement

On arrival in a testing room, all participants in the 3 studies first completed the Positive and Negative Affect Scale (PANAS²) and the Life Orientation Test Revised scale (LOT-R³) to measure their mood and optimistic trait. PANAS was administered again after the experiment to monitor their mood change. In Studies 2 and 3, participants also completed the Beck Depression Inventory (BDI⁴), the Dysfunctional Attitude Scale (DAS⁵) and the State Trait Anxiety Inventory (STAI⁶) before IN-OT/PL. The BDI, a 21-item multiple-choice inventory, was employed to measure depressive symptoms. Participants' cognitive distortions were measured using the 40-item DAS, which was designed to identify and measure cognitive distortions related to depression. Lower scores on DAS represent more adaptive beliefs and fewer cognitive distortions. Participant's trait and state anxiety was measured using the STAI, which contains 20 items for assessing trait anxiety and 20 for state anxiety. All items were rated on a 4-point scale, with higher scores indicating greater anxiety. After the experiment, PANAS was administered again to monitor mood change.

Data analysis

Hierarchical regression analyses. We performed hierarchical regression analyses to assess whether individual differences in depression or anxiety traits moderated OT effects on belief update (BU). We normalized the independent variable (Treatment, coded as a dichotomous dummy variable in which 0 represented PL and 1 represented IN-OT) and the covariate variable (normalized BDI, DAS and TA scores, respectively). Three moderated hierarchical regression models were built, respectively with BDI, DAS, or TA

scores as moderator. For each model, normalized Treatment, BDI, DAS, or TA scores, and their interaction were sequentially entered as predictor variables. These analyses were conducted separately with BU_{Des} and BU_{Undes} as dependent variable. The significant Treatment x Trait interaction was followed up with tests of simple slopes, which assessed the magnitude of different effects that contributed to an interaction.

Learning rate. Learning rate was calculated as the strength of the association between the estimation error (prediction error, PE) and the subsequent updates (update) for desirable and undesirable trials, respectively. The learning rate has been suggested as a computational principle that underlies the observed biased belief formation by pointing to estimation errors as a learning signal⁷ and reflects the dynamic learning processes of positive and negative prediction errors⁸. We made a linear regression of participant's updates as a function of estimation errors. The learning rate (the slope of this linear regression, β) indicates how well a person integrates good and bad news into beliefs. The larger the β the more participants rely on estimation errors to form a new estimate. BU_{Des} and BU_{Undes} were separately regressed onto PEs, resulting in two standardized regression coefficient: β_{Des} and β_{Undes} . We then examined OT effects on learning rate to determine how OT influenced learning from desirable and undesirable feedback. To do so, learning rates (β) were transformed to Z scores using Fisher's transformation: $Z = \frac{1}{2} \ln\left(\frac{1+\beta}{1-\beta}\right)$, and subjected to Treatment x Feedback ANOVAs.

Mediation analysis. We performed mediation analyses to examine whether the effects of OT on the optimistic bias (OB, indexed by BU_{Des} minus BU_{Undes}) occurred through the OT effects on confidence update or acceptance of feedback. Similar to our previous studies⁹, a bootstrapping method was used to estimate the mediation effect. Bootstrapping is a nonparametric approach to effect-size estimation and hypothesis testing that is increasingly recommended for many types of analyses, including mediation^{10,11}. Rather than imposing questionable distributional assumptions, bootstrapping generates an empirical approximation of the sampling distribution of a statistic by repeated random resampling from the available data, and uses this distribution to calculate p-values and construct confidence intervals (5,000 resamples were taken for each effect).

Supporting figures

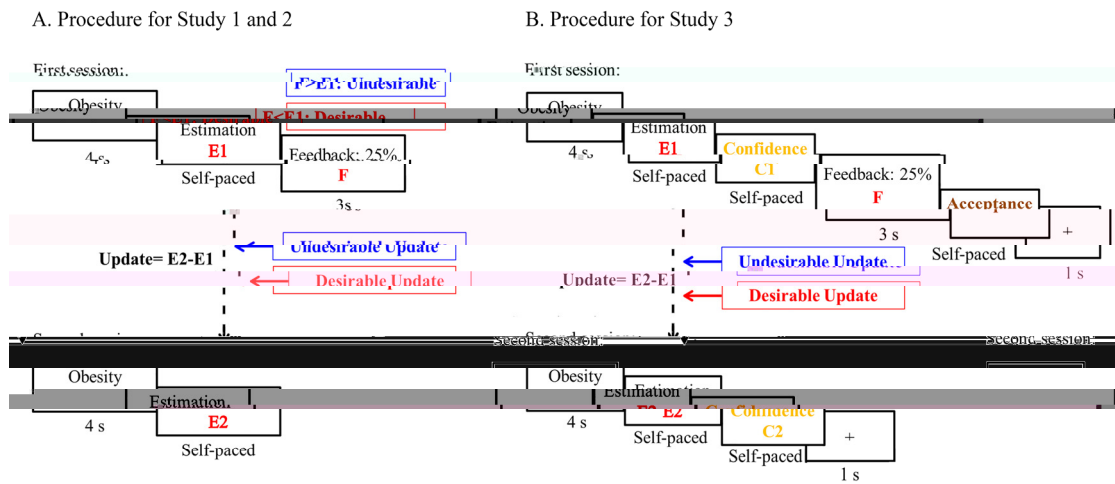


Fig. S1. Illustration of experimental procedures in the current work. In Study 1 (discovery sample) and Study 2 (replication sample), participants completed two sessions of adverse life event estimation (A). In the first session participants were presented with 40 different adverse life events and had to estimate their likelihood of experiencing each life event on a self-paced basis (1st estimation). Participants were then presented with the probability of each event occurring to an average people in a similar socio-cultural environment (feedback). In the second session, participants were presented with the 40 adverse life events in a random order and had to estimate the likelihood of each event again in (2nd estimation). The belief update task in Study 3 was similar to that in Studies 1 and 2, except that, for each event, participants were asked to rate 1) their confidence of the 1st and 2nd Estimate (ranging from 0% to 99%) after their estimation; and 2) their acceptance of the feedback (ranging from 0% to 99%) after the presentation of the feedback probability (B).

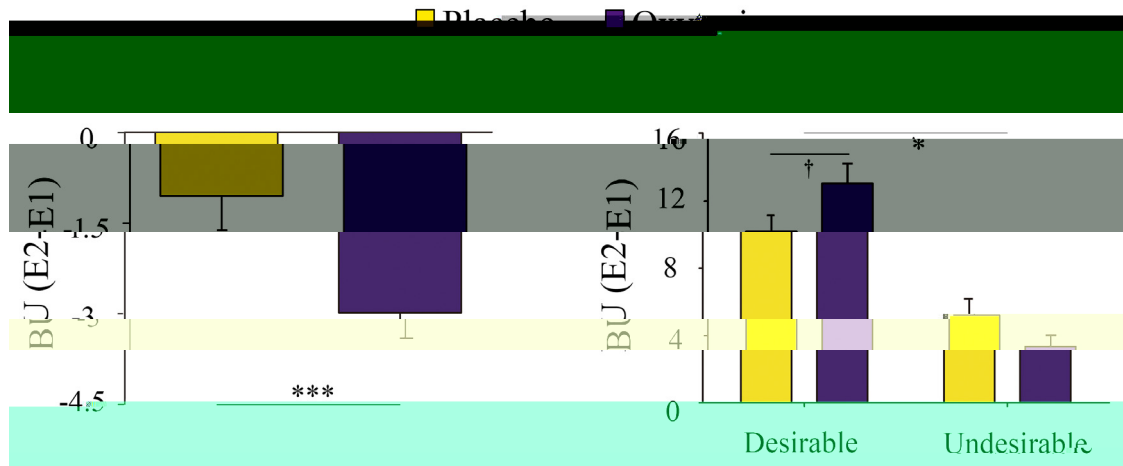


Fig. S2. Distinct OT effects on belief updates in response to desirable and undesirable feedback in Study 3. IN-OT enhanced belief updating upon desirable feedback, but decreased belief updating upon undesirable feedback (***) $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$).

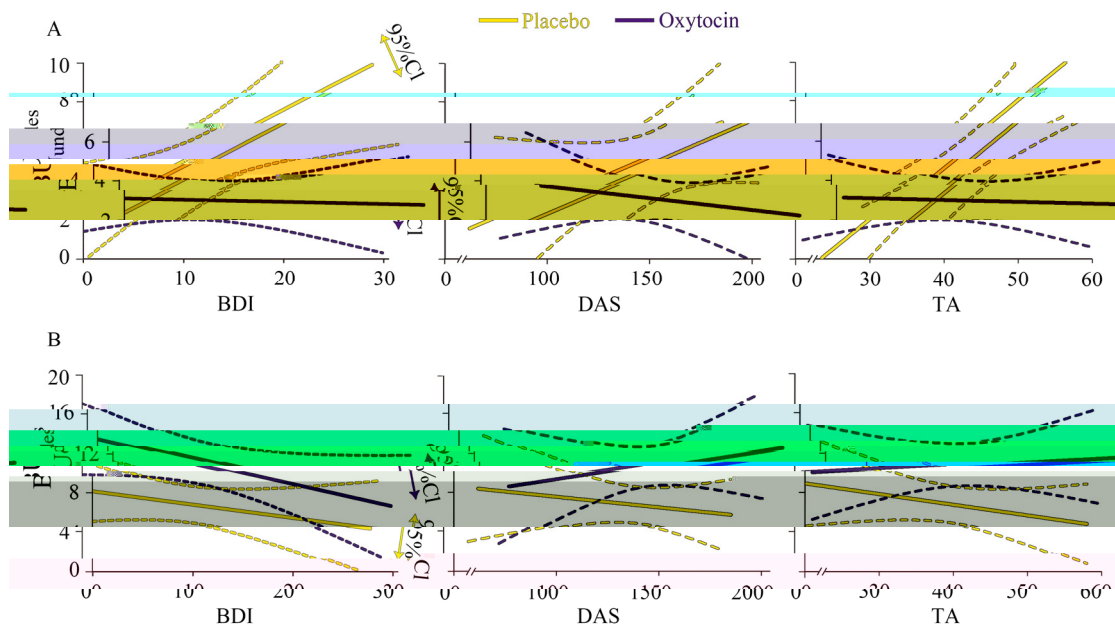


Fig. S3. The results of Treatment x Trait interaction on belief updating in Study 2.

Treatment x Trait interaction predicted belief updating upon undesirable feedback (A), but not upon desirable feedback (B) in Study 2. BDI = Beck's depression inventory; DAS= Dysfunctional Attitude Scale; TA = Trait Anxiety.

The moderated hierarchical regression models regressed the moderator (normalized BDI, DAS and TA scores, respectively), independent variable (Treatment), and their interactions onto BU_{Des} and BU_{Undes} , respectively. The analyses of Study 2 showed that the interaction between Treatment and Trait was predictive of BU_{Undes} (BDI: $B = -0.41$, $t(80) = -2.48$, $p=0.015$; DAS: $B = -0.27$, $t(80) = -1.72$, $p=0.089$; TA: $B = -0.57$, $t(80) = -3.74$, $p<0.001$, Fig. S3A; Table S3-5); but not BU_{Des} (BDI: $B = -0.08$, $t(80) = -0.51$, $p=0.613$; DAS: $B = 0.15$, $t(80) = 0.99$, $p=0.327$; TA: $B = 0.16$, $t(80) = 0.97$, $p=0.335$, Fig. S3B; Table S3-5), suggesting that individuals' depression and anxiety traits moderated OT effects on belief updates in response to undesirable feedback.

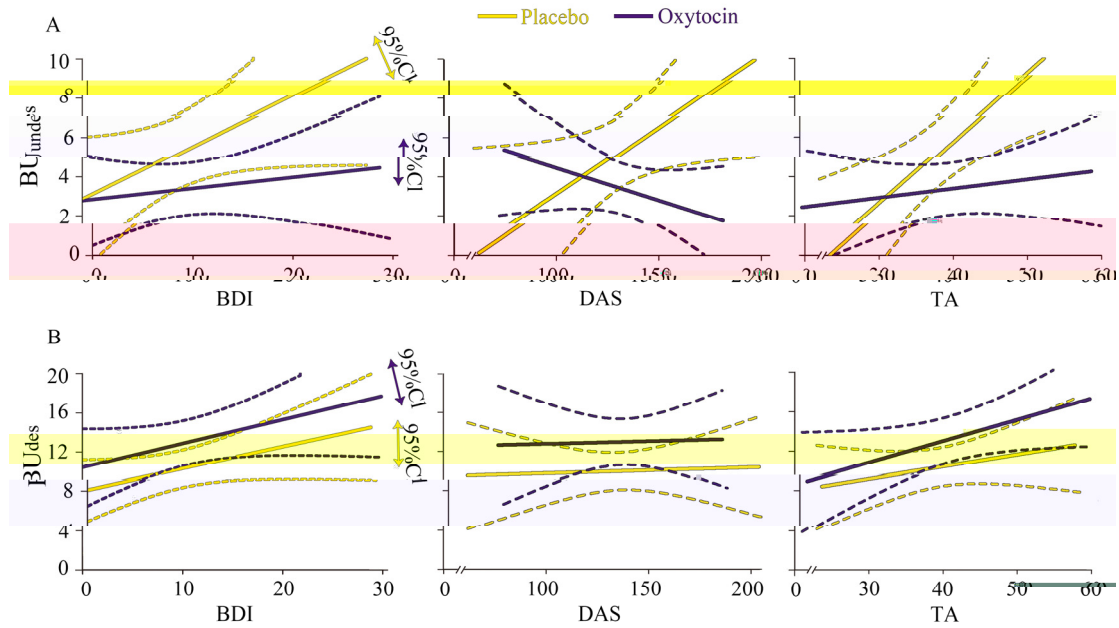


Fig. 4. The results of Treatment x Trait interaction on belief updating in Study 3.

Treatment x Trait interaction predicted belief updating upon undesirable feedback (A), but not upon desirable feedback (B) in Study 3. BDI = Beck's depression inventory; DAS= Dysfunctional Attitude Scale; TA = Trait Anxiety.

The moderated hierarchical regression models regressed the moderator (normalized BDI, DAS and TA scores, respectively), independent variable (Treatment), and their interactions onto BU_{Des} and BU_{Undes} , respectively. The analyses of Study 3 showed that the interaction between Treatment and Trait was predictive of BU_{Undes} (BDI: $B = -0.17$, $t(110) = -1.24$, $p=0.218$; DAS: $B = -0.30$, $t(109) = -2.41$, $p=0.018$; TA: $B = -0.33$, $t(110) = -2.33$, $p=0.022$, Fig. S4A; Table S3-5); but not BU_{Des} (BDI: $B = 0.01$, $t(110) = 0.10$, $p=0.917$; DAS: $B = -0.001$, $t(109) = -0.01$, $p=0.991$; TA: $B = 0.09$, $t(110) = 0.58$, $p=0.562$, Fig. S4B; Table S3-5), suggesting that individuals' depression and anxiety traits moderated OT effects on belief updates in response to undesirable feedback. Note: The Treatment x BDI interaction on undesirable updating was reliable in Study 2, and when combined data of Studies 2 and 3. This effect did not reach significant in Study 3 but showed the same pattern as that in Study 2 and combined dataset.

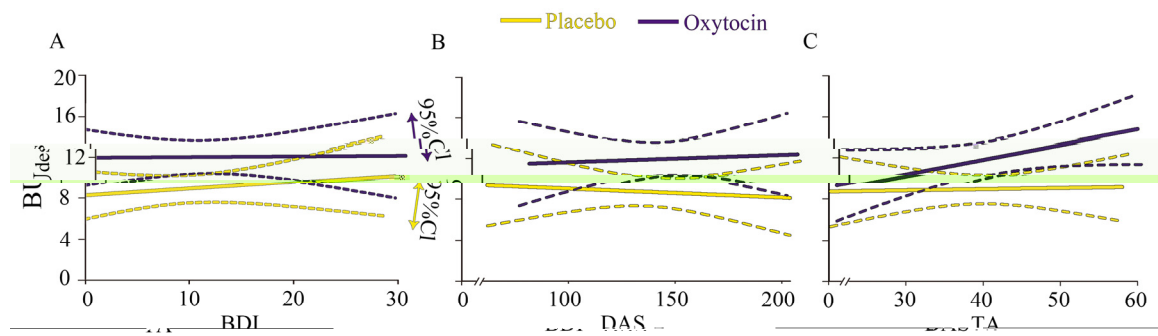


Fig. S5. The results of Treatment x Trait interaction on belief updating upon desirable feedback in data collapsed over Studies 2 and 3. There was no significant Treatment x Trait interaction on belief updating upon desirable feedback (BDI: $B = -0.045$, $t(194) = -0.42$, $p = 0.677$, DAS: $B = 0.040$, $t(193) = 0.40$, $p = 0.690$; TA: $B = 0.123$, $t(194) = 1.12$, $p = 0.265$). BDI = Beck's depression inventory; DAS = Dysfunctional Attitude Scale; TA = Trait Anxiety.

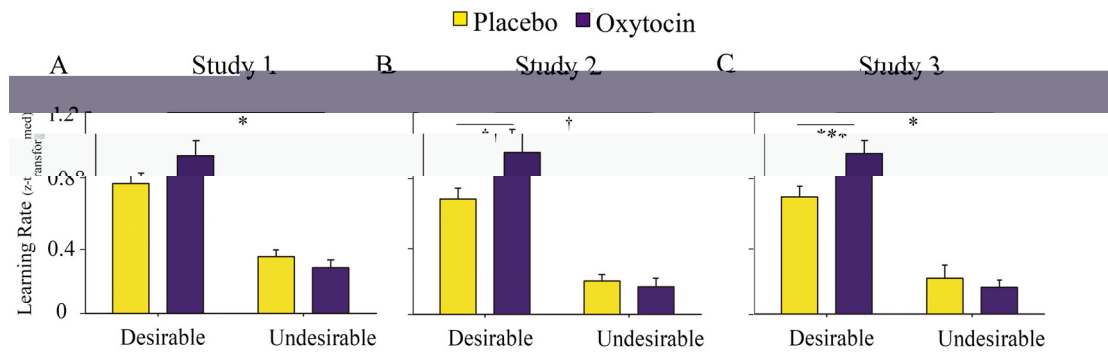
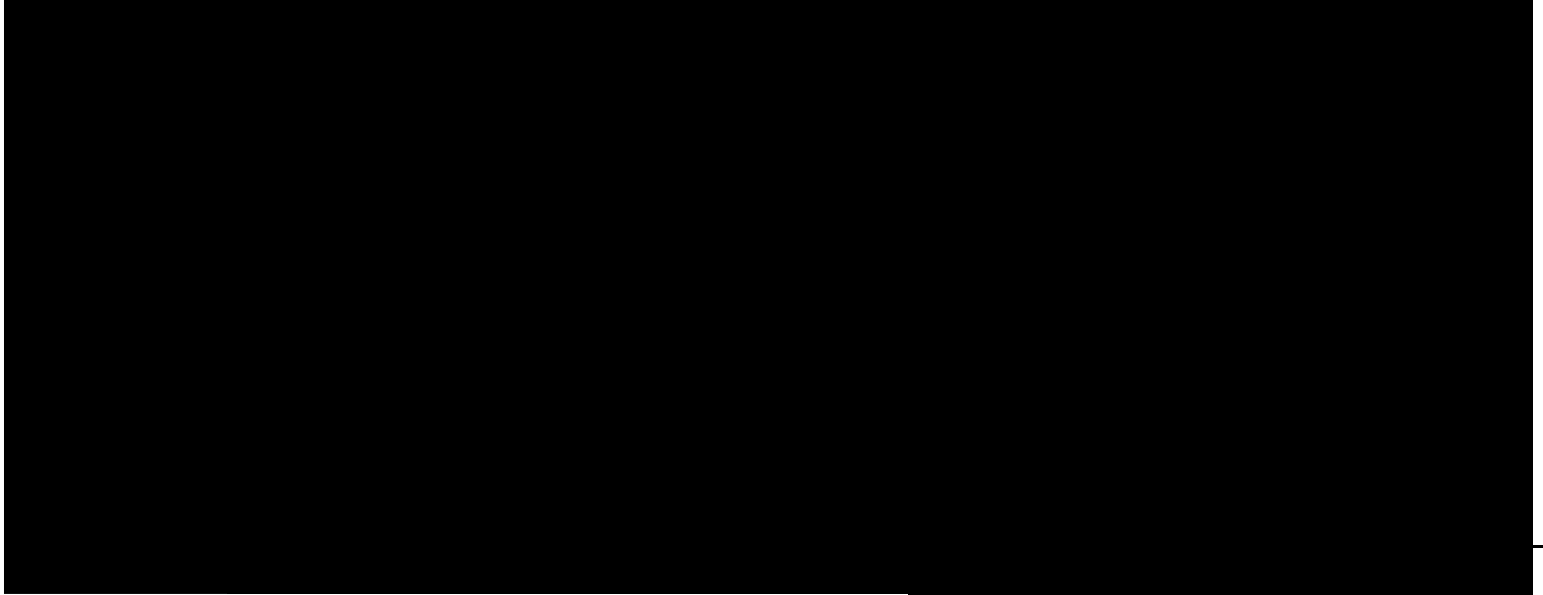


Fig. S6. OT effects on the learning rate for each study. OT, compared to PL, enhanced the strength of the association between estimation error and subsequent update in response to desirable feedback not undesirable feedback in each study.

We found that participants learned to a greater degree from estimation errors in the desirable (than undesirable) trials (Study 1: $F(1, 97)= 89.252, p<0.001, \eta^2=0.479$; Study 2: $F(1, 93)= 64.647, p<0.001, \eta^2=0.410$; Study 3: $F(1, 112)= 97.512, p<0.001, \eta^2=0.465$). Moreover, a significant Treatment x Feedback interaction on the learning rate confirmed that the OT selectively increased participants' learning from prediction error in the desirable but not undesirable trials (Study 1: $F(1, 97)= 3.989, p=0.049, \eta^2=0.039$; Study 2: $F(1, 93)= 3.842, p=.053, \eta^2=0.040$; Study 3: $F(1, 112)= 5.894, p=0.017, \eta^2=0.050$).



	OT	-3.31(7.13)	13.11(10.16)	3.75 (4.77)			
Study 2	PL	-0.36(5.06)	7.83(6.03)	4.93(4.64)	-0.63 (0.23)	0.68(0.44)	0.20(0.26)

Table S2. Self-reports of adverse life events characteristics

Variables	Study 2			Study 3		
	PL M (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)
Familiarity	3.70 (1.21)	3.56 (0.70)	0.62 (0.54)	3.69(1.12)	3.55(0.87)	0.72(0.48)
Negativity	4.30 (0.78)	4.07 (0.67)	1.49 (0.14)	4.08(0.90)	4.20(0.75)	-0.75(0.45)
Vividness	3.98 (1.05)	3.80 (0.89)	0.85 (0.40)	3.90(0.91)	3.84(1.00)	0.37(0.72)
Arousal	3.86 (0.86)	3.81 (0.72)	0.31 (0.76)	3.73(0.85)	3.83(0.73)	-0.68(0.50)
Prior experience	1.22 (0.20)	1.24 (0.27)	-0.40 (0.69)	1.23(0.19)	1.24(0.29)	-0.34(0.74)

The rating scores of familiarity, negativity, vividness, arousal and prior experience for adverse life events (on 7-point scales: 1=not familiar/negative/vivid/aroused at all; never occurred to me; 7=extremely familiar/negative/vivid/aroused; frequently occurred to me) were compared between OT and PL groups as manipulation check of whether the characteristics of adverse life events were similar between the PL and OT groups. There was no group difference in Studies 2 or 3, for familiarity, negativity, vividness, arousal and prior experience ratings.

Table S3. The results of the hierarchical regression analyses on Update_{Undesirable} with BDI scores as moderator in Study 2 and Study 3, respectively.

Predictors	Study 2				Study 3			
	BU _{Undes}		BU _{Des}		BU _{Undes}		BU _{Des}	
	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2
Step 1								
Treatment	-0.25*	0.090*	0.31**	0.133**	-0.16†	0.052†	0.16†	0.076*
BDI	0.19		-0.22*		0.17†		0.22*	
Step 2								
Treatment	-0.41*	0.065*	-0.08	0.003	-0.17	0.013	0.01	0.001
×BDI								
Total (R^2)		0.155**		0.136**		0.065†		0.076*

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$;

BDI: Participant's scores in Beck Depression Inventory.

In the regression analyses, dummy coded Treatment variable and standardized continuous BDI (or DAS, TA in the following tables) scores were entered in step 1 regression; Treatment \times BDI (or Treatment \times DAS, Treatment \times TA) were entered in step 2 to predict desirable or undesirable update as dependent variables separately.

Table S4. The results of the hierarchical regression analyses on Update Undesirable with DAS scores as moderator in Study 2 and Study 3, respectively.

Predictors	Study 2				Study 3			
	BU _{Undes}		BU _{Des}		BU _{Undes}		BU _{Des}	
	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2
Step 1								
Treatment	-0.25*	0.065†	0.29**	0.085*	-0.16†	0.032	0.19†	0.035
DAS	0.09		0.03		0.10		0.02	
Step 2								
Treatment	-0.27†	0.033†	0.15	0.011	-0.30*	0.049*	-0.001	0.001
×DAS								
Total (R^2)		0.098*		0.096*		0.081*		0.035
N		83		83		113		113

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$;

DAS: Participant's scores in Dysfunctional Attitude Scale.

Table S5. The results of the hierarchical regression analyses on Update Undesirable with TA scores as moderator in Study 2 and Study 3, respectively.

Predictors	Study 2 (Replication Study)				Study 3			
	BU _{Undes}		BU _{Des}		BU _{Undes}		BU _{Des}	
	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2
Step 1								
Treatment	-0.25*	0.131**	0.29**	0.085*	-0.17†	0.079*	0.15	0.072*
TA	0.27**		-0.03		0.24**		0.21*	
Step 2								
Treatment	-0.57***	0.130***	0.16	0.010	-0.33*	0.043*	0.09	0.003
×TA								
Total (R^2)		0.261***		0.095*		0.122**		0.075*
N		83		83		113		113

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$;

TA: Participant's scores in Trait Anxiety.

Table S6. The results of simple slope analysis (breaking down the Treatment x Trait interaction by analyzing OT effect for less and well socially adapted individuals)

Slope for individuals with low trait scores		
	Study 2	Study 3
BDI	b =-0.014, t(80) =-0.011, p=0.991	b =-0.571, t(110) =-0.347, p=0.729
DAS	b =-0.547, t(80) =-0.436, p=0.664	b =0.840, t(109) =0.510, p=0.611
TA	b =0.972, t(80) =0.849, p=0.399	b =0.519, t(110) =0.325, p=0.746

Slope for individuals with high trait scores		
	Study 2	Study 3
BDI	b =-4.386, t(80) =-3.489, p=0.001	b =-3.471, t(110) =-2.100, p=0.038
DAS	b =-3.619, t(80) =-2.836, p=0.006	b =-4.785, t(109) =-2.914, p=0.004
TA	b =-5.172, t(80) =-4.466, p<0.001	b =-4.869, t(110) =-2.986, p=0.003

Table S7. The results of simple slope analysis (breaking down the Treatment x Trait interaction by analyzing trait effects on belief updating under OT and placebo, respectively)

Slope for PL group		
	Study 2	Study 3
BDI	b =2.098, t(80) =3.055, p=0.003	b =1.869, t(110) =2.184, p=0.031
DAS	b =1.209, t(80) =1.845, p=0.069	b =1.911, t(109) =2.399, p=0.018
TA	b =2.983, t(80) =4.698, p<0.001	b =3.139, t(110) =3.491, p=0.001

Slope for OT group		
	Study 2	Study 3
BDI	b =-0.088, t(80) =-0.158, p=0.875	b =0.419, t(110) =0.525, p=0.600
DAS	b =-0.327, t(80) =-0.539, p=0.592	b =-0.901, t(109) =-1.057, p=0.293
TA	b =-0.089, t(80) =-0.172, p=0.864	b =0.445, t(110) =0.611, p=0.542

Table S8 Information of the three scales used in the current study (data collapsed over Studies 2 and 3)

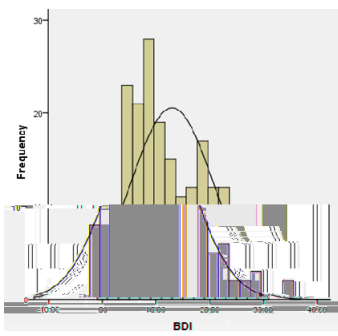
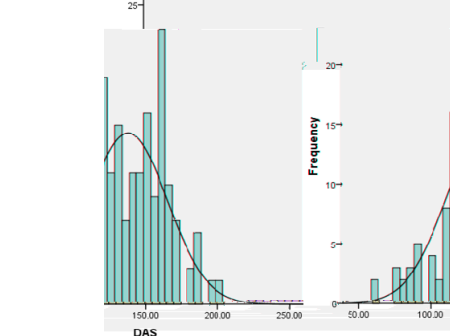
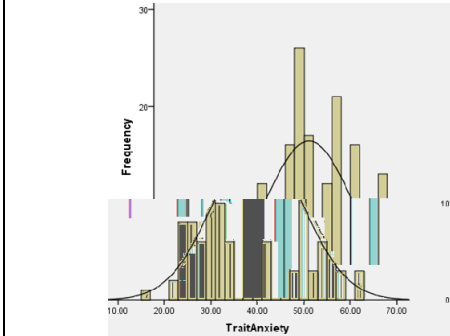
Scales	Beck Depression Inventory (BDI)	Dysfunctional Attitude Scale (DAS)	State-Trait Anxiety Inventory-Trait Anxiety (TA)
Description	BDI ⁴ is a 21-item self-report inventory with excellent test-retest reliability and validity. It measures depression severity in not only clinical patients but also college populations ¹⁷ .	DAS ⁵ is a 40-item scale, designed to measure cognitive distortions related to depression, with good-to-excellent levels of test-retest reliability, and criterion validity ¹⁸ .	TA ⁶ is a 20 item scale assessing trait anxiety, with good internal consistency, test-retest reliability, discriminating anxiety disorders from healthy controls ¹⁹ .
Mean (SD)	10.44 (7.67); comparable to previous study of 9.14(8.45) in 15,233 college students ¹²⁰ .	138.05(27.36); similar to previous study of 137.8 (23.6) in large community sample of 8,960 adults ²¹ .	40.23(9.97); similar to that obtained in the original STAI manual (M = 39.6, SD = 9.79 ⁶).
Range	0-35	62-204	16-62
Distribution			
Scale reliability	0.878 (Similar to that given in the BDI studies meta-analysis; $r=0.84^{22}$).	0.903 (Similar to that given in previous studies, $r = 0.85^{231}$; $r=0.86^{21}$).	0.913 (Similar to that given in the original manual: $r=0.90^6$).
Discriminant validity	BDI & DAS: $\chi^2 (3) = 449.60, p<0.001$; DAS & TA: $\chi^2 (3) = 75.58, p<0.001$; BDI & TA: $\chi^2 (3) = 327.61, p<0.001$.		

Table S9 Hierarchical regression analyses on belief updates upon desirable and undesirable feedback with BDI as moderator (data collapsed over Studies 2 and 3)

Predictors	BU _{Undes}		BU _{Des}	
	β	ΔR^2	β	ΔR^2
Step 1 (enter)				
Treatment	-0.19**	0.060**	0.20	0.042*
BDI	0.17*		0.03	
Step 2 (enter)				
Treatment x BDI	-0.25*	0.026*	-0.05	0.001
Total (R^2)		0.086***		0.043*
N		197		197

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$;

BDI: Participant's scores in Beck Depression Inventory.

In the regression analyses, dummy coded Treatment variable and standardized continuous BDI (or DAS, TA in the following tables) scores were entered in step 1 regression; Treatment \times BDI (or Treatment \times DAS, Treatment \times TA in the following tables) were entered in step 2 to predict BU_{Des} or BU_{Undes} as dependent variables separately.

Table S10. Hierarchical regression analyses on belief updates upon desirable and undesirable feedback with DAS as moderator (data collapsed over Studies 2 and 3)

Predictors	BU _{Undes}		BU _{Des}	
	β	ΔR^2	β	ΔR^2
Step 1 (enter)				
Treatment	-0.19**	0.040*	0.21**	0.045*
DAS	0.09		-0.002	
Step 2 (enter)				
Treatment x DAS	-0.29**	0.041**	0.04	0.001
Total (R^2)		0.082***		0.045*
N		196		196

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$;

DAS: Participant's scores in Dysfunctional Attitude Scale.

The hierarchical regression analysis revealed a significant Treatment \times DAS interaction on BU_{Undes} but not BU_{Des}.

Table S11. Hierarchical regression analyses on belief updates upon desirable and undesirable feedback with TA as moderator (data collapsed over Studies 2 and 3)

Predictors	BU _{Undes}		BU _{Des}	
	β	ΔR^2	β	ΔR^2
Step 1 (enter)				
Treatment	-0.19**	0.091***	0.19**	0.053**
TA	0.25***		0.11	
Step 2 (enter)				
Treatment \times TA	-0.40***	0.063***	0.12	0.006
Total (R^2)		0.154***		0.059**
N		197		197

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$;

TA: Participant's scores in Trait Anxiety.

The hierarchical regression analysis revealed a significant Treatment \times TA interaction on BU_{Undes} but not BU_{Des}.

Table S12. The results of mediation analysis to test OT effect on confidence update upon desirable feedback (CU_{Des}) as a mediator of its effect on optimistic bias (OB, indexed by $BU_{Des} - BU_{Undes}$).

Variable	<i>Coeff</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Regression Model 1 (Total effect of Treatment on OB)				
Treatment	4.67*	1.90	2.46	0.016
Dependent: OB				
Regression Model 2 (Treatment to CU_{Des})				
Independent: Treatment	5.63***	1.54	3.64	0.0004
Mediator: CU_{Des}				
Direct effects of mediator on OB				
Independent: Treatment	0.34**	0.11	3.02	0.003
Remaining direct effect of Treatment on OB				
Independent: Treatment	2.76	1.94	1.42	0.157
Indirect effect of Treatment on OB via CU_{Des} (Sobel test result)				
CU_{Des}	1.91*	0.84	2.36	0.018
	<i>Coeff</i>	<i>SE</i>	<i>LLCI95</i>	<i>ULCI95</i>
Indirect effect of Treatment on OB via CU_{Des} (bootstrap results)				
CU_{Des}	1.91*	0.88	0.58	4.32

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. Confidence intervals for indirect effect are bias-corrected and accelerated; bootstrap resamples=5000; $N=114$ for all tests.

Table S13. The results of mediation analysis to test OT effect on confidence update upon desirable feedback (CU_{Undes}) as a mediator of its effect on optimistic bias (OB, indexed by $BU_{Des} - BU_{Undes}$).

Variable	<i>Coeff</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Regression Model 1 (Total effect of Treatment on OB)				
Treatment	4.67*	1.90	2.46	0.016

Table S14. The results of mediation analysis to test OT effect on acceptance of desirable feedback (AC_{Des}) as a mediator of its effect on optimistic bias (OB, indexed by $BU_{Des} - BU_{Undes}$).

Variable	<i>Coeff</i>	<i>SE</i>	<i>t</i>	<i>p</i>
-----------------	--------------	-----------	----------	----------

Table S15. The results of mediation analysis to test OT effect on acceptance of desirable feedback (AC_{Undes}) as a mediator of its effect on optimistic bias (OB, indexed by $BU_{Des} - BU_{Undes}$).

Table S16. Participant information for each study

Variable	Study 1			Study 2			Study 3		
	PM (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)
Num.	50	49	—	47	48	—	57	57	—
Age	22.89(3.01)	22.03(2.55)	1.38 (0.17)	22.43(2.32)	22.94(2.22)	-1.10(0.27)	22.70(2.51)	22.54(2.11)	0.36(0.72)
LOT-R	22.29(3.27)	22.03(3.05)	0.37 (0.71)	22.69(2.79)	22.81(2.86)	-0.21(0.83)	22.89(3.23)	22.56(2.95)	0.58(0.57)

Note:

LOT-R: Participants' scores in Life Orientation Test-Revised.

For the demographic variables (age) and life orientation scores, there is no significant difference between OT and PL groups in each of the three studies.

Table S17. Questionnaire measures in Studies 2 and 3.

Variables	Study 2			Study 3		
	PL M (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)
BDI	9.94 (7.75)	11.38 (8.28)	-0.81 (0.423)	9.47 (7.18)	10.58 (7.71)	-0.79(0.430)
DAS	138.89 (28.39)	143.31 (26.41)	-0.74 (0.464)	134.93(28.09)	138.47(27.03)	-0.68(0.496)
TA	39.81(9.61)	40.44 (10.13)	-0.29 (0.773)	39.14(8.47)	40.86(10.46)	-0.96(0.337)
SA	35.86 (10.12)	35.02 (9.17)	0.40 (0.692)	34.79(8.20)	35.77(9.67)	-0.59(0.560)

Note:

BDI : Participants' scores in Beck Depression Inventory; DAS: Participants' scores in Dysfunctional Attitude Scale; TA: Participants' scores in Trait Anxiety; SA: Participants' scores in State Anxiety.

The Independent Samples t-test was employed to compare the scores of BDI, DAS, TA, SA between the OT and PL groups in Study 2 and Study 3, respectively. There was no group difference on the BDI, DAS, TA and SA scores in Study 2 or 3.

Table S18. Mood changes from pre-experiment to post-experiment for each study

Mood	Study 1			Study 2			Study 3		
	PL M (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)
Pre-positive	32.53 (7.86)	31.86 (6.75)	0.41 (0.685)	31.07 (5.79)	31.96 (6.44)	-0.70 (0.486)	31.44 (6.09)	31.04 (6.74)	0.40 (0.691)
Pre-negative	16.86 (6.70)	15.92 (7.04)	0.62 (0.539)	16.33 (6.70)	16.81 (7.03)	-0.34 (0.738)	16.12 (7.17)	16.39 (5.77)	0.18 (0.857)
Post-positive	31.75 (8.59)	32.81 (8.99)	-0.54 (0.589)	31.73 (7.45)	31.48 (7.74)	0.16 (0.872)	32.53 (7.60)	30.88 (7.41)	0.73 (0.465)
Post-negative	16.95 (6.39)	15.30 (6.19)	1.18 (0.242)	15.95 (5.92)	16.70 (7.10)	-0.54 (0.588)	16.09 (5.86)	16.88 (7.02)	-0.10 (0.919)
positive	-0.73 (6.63)	-0.02 (6.01)	-0.50 (0.622)	0.15 (2.79)	-0.05 (0.57)	0.47 (0.637)	0.11 (0.71)	-0.02 (0.59)	0.49 (0.625)
negative	0.09 (4.71)	-1.59 (6.69)	1.33 (0.189)	-0.16 (0.99)	-0.02 (0.54)	-0.84 (0.401)	-0.01 (0.47)	0.04 (0.60)	-0.33 (0.741)

Note:

positive= Post-positive – Pre-positive; negative= Post- negative – Pre- negative.

OT and PL groups did not differ in mood both before and after the treatment. Moreover, participant’s mood change before and after treatment was not different between OT and PL groups in each of the three studies

Table S19. Memory error (%) for feedback in each study.

Study	Groups	Total	Desirable trials	Undesirable trials
Study 1	PL: M (SD)	2.22 (4.59)	4.92 (5.20)	0.59 (5.80)
	OT: M (SD)	0.75 (5.27)	4.54 (7.43)	2.16 (6.57)
	PL vs. OT: F(p)	0.03(0.854)	0.754(0.388)	0.28(0.597)
Study 2	PL: M (SD)	1.48 (4.51)	5.35 (6.61)	-0.90 (4.55)
	OT: M (SD)	0.17 (3.73)	3.55 (5.50)	-3.02 (4.31)
	PL vs. OT: F(p)	0.27(0.604)	0.19(0.661)	1.73(0.192)
Study 3	PL: M (SD)	1.57 (4.05)	4.89(6.14)	-0.96(4.87)
	OT: M (SD)	1.38 (4.52)	5.68 (6.76)	-1.43 (4.78)
	PL vs. OT: F(p)	0.03(0.862)	0.508(0.478)	0.002(0.967)

The difference between recalled feedback and actually presented feedback was used to indicate memory performance of feedback (Memory error). We compared memory errors respectively for all trials, desirable trials and undesirable trials between the OT and PL groups to see whether OT affected the memory of feedback in each of the three studies. ANCOVA F-test with participants' own estimates as covariate variables has not found consistent significant difference between OT and PL groups in different conditions.

Table S20. Reaction times (RTs, ms) for 1st and 2nd estimation in each study

Study	Groups	1 st estimation	2 nd Estimates		2 nd Estimates
				(Desirable trials)	(Undesirable trials)
Study 1	PL: M (SD)	2973.59(870.80)	2021.31(621.03)	1897.96(747.67)	1856.55(689.24)
	OT: M (SD)	2742.68(835.89)	1959.32(717.35)	1833.53(704.30)	1781.11(868.35)
	PL vs. OT: T (p)	1.34(0.184)	0.46(0.648)	0.44(0.662)	0.48(0.636)
Study 2	PL: M (SD)	2496.48(901.90)	1781.40(521.27)	1760.93(630.59)	1683.30(524.30)
	OT: M (SD)	2538.54(929.57)	1984.31(688.36)	1985.99(758.20)	1859.53(763.96)
	PL vs. OT: T (p)	-0.21(0.833)	-1.49(0.141)	-1.45(0.150)	-1.20(0.234)
Study 3	PL: M (SD)	1831.50(516.12)	1558.26(584.68)	1561.14(595.61)	1497.93(463.37)
	OT: M (SD)	1873.83(707.45)	1561.19(553.52)	1487.43(557.18)	1546.59(567.62)
	PL vs. OT: T (p)	-0.36(0.716)	-0.03(0.978)	0.68(0.496)	-0.50(0.617)

Table S21. Mean (SDs) number of desirable and undesirable trials for each study.

Study		Desirable trials	Undesirable trials
Study 1	PL: M (SD)	15.38(5.39)	23.10(5.43)
	OT: M (SD)	15.59(7.20)	22.90(7.09)
	PL vs. OT: T (p)	-0.17(0.869)	0.16(0.874)
	ANOVA	Treatment x Feedback Interaction: F (1, 97)=0.027, p=0.870	
Study 2	PL: M (SD)	14.94(6.03)	23.79(6.33)
	OT: M (SD)	15.42(6.73)	23.35(6.82)
	PL vs. OT: T (p)	-0.37(0.715)	0.32(0.749)
	ANOVA	Treatment x Feedback Interaction: F (1, 93)=0.12, p=0.732	
Study 3	PL: M (SD)	16.28(8.03)	21.98(8.22)
	OT: M (SD)	14.86(7.35)	23.68(7.34)
	PL vs. OT: T (p)	0.99(0.327)	-1.17(0.246)
	ANOVA	Treatment x Feedback Interaction: F (1, 112)=1.17, p=0.282	

Reference

1. Sharot T, Korn CW, Dolan RJ (2011) How unrealistic optimism is maintained in the face of reality. *Nat Neurosci* 14(11):1475-1479.
2. Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol* 54(6): 1063-1070.
3. Scheier MF, Carver CS, Bridges MW (1994) Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *J Pers Soc Psychol* 67(6):1063-1078.
4. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961) An inventory for measuring depression. *Arch Gen Psychiatry* 4(6):561-571.
5. Weissman AN, Beck AT (1978) *Development and Validation of the Dysfunctional Attitude Scale: A Preliminary Investigation*. Paper presented at the

Annual Meeting of The Association for the Advancement of Behavior Therapy, Chicago.

6. Spielberger CD, Gorsuch RL (1983) *State-trait anxiety inventory for adults: Manual, instrument, and scoring guide*. (Mind Garden, Incorporated).
7. Garrett N, et al. (2014) Losing the rose tinted glasses: neural substrates of unbiased belief updating in depression. *Front Hum Neurosci* 8:639.
8. Palminteri S, et al. (2012) Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron* 76(5):998–1009.
9. Ma Y, Liu Y, Rand DG, Heatherton TF, Han S (2015) Opposing Oxytocin Effects on Inter-Group Cooperative Behavior in Intuitive and Reflective Minds. *Neuropsychopharmacol* 40(10):2379-2387.
10. Mackinnon DP, Lockwood CM, Williams J (2004) Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivar Behav Res* 39(1):99-128.
11. Shrouf PE & Bolger N (2002) Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol Methods* 7(4):422-445.
12. Preacher KJ, Rucker DD, Hayes AF (2007) Assessing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions. *Multivar Behav Res* 42(1):185-227.
13. Preacher KJ, Hayes AF (2008) Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods* 40(3):879-891.
14. Hayes AF (2013) *Introduction to mediation, moderation, and conditional process analysis* (The Guilford Press, New York).
15. Baron RM, Kenny DA (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 51(6):1173-1182.
16. Sobel ME, Sobel ME (1982) Asymptotic intervals for indirect effects in structural equations models. *Sociol Methodol* 13(13):290-312.

17. Dobson KS, Breiter HJ (1983) Cognitive assessment of depression: reliability and validity of three measures. *J Abnorm Psychol* 92(1):107-109.
18. Spielberger CD, Gorsuch RL (1983) *State-trait anxiety inventory for adults: Manual, instrument, and scoring guide*. (Mind Garden, Incorporated).
19. Spielberger CD, Reheiser EC (2009) Assessment of Emotions: Anxiety, Anger, Depression, and Curiosity. *Appl Psychol-Hlth We* 1(3):271–302.
20. Whisman MA, Richardson ED (2015) Normative Data on the Beck Depression Inventory – Second Edition (BDI-II) in College Students. *J Clin Psychol* 71(9):898–907.
21. Graaf LED, Roelofs J, Huibers MJH (2009) Measuring Dysfunctional Attitudes in the General Population: The Dysfunctional Attitude Scale (form A) Revised. *Cognitive Ther Res* 33(4):345-355.
22. Beck AT, Steer RA (1984) Internal consistencies of the original and revised Beck Depression Inventory. *J Clin Psychol* 40(6):1365-1367.
23. Oliver JM, Baumgart EP (1985) The Dysfunctional Attitude Scale: Psychometric properties in an unselected adult population. *Cognitive Ther Res* 9(2):161-167.