NEURAL CORRELATES OF CONDITIONAL AND UNCONDITIONAL TRUST IN TWO-PERSON RECIPROCAL EXCHANGE

Frank Krueger[1], Kevin McCabe[2], Jorge Moll[3], Nikolaus Kriegeskorte[4], Roland Zahn[1], Maren Strenziok[1], Armin Heinecke[5], Jordan Grafman[1]

[1] *Cognitive Neuroscience Section, NINDS, National Institutes of Health, Bethesda, USA*

[2] *The Center for the Study of Neuroeconomics, George Mason University, USA*

[3] *Cognitive and Behavioral Neuroscience Unit, LABS–D'Or Hospital Network, Rio de Janeiro, Brazil*

[4] *Laboratory of Brain and Cognition, NIMH, National Institutes of Health, Bethesda, USA*

[5] *Brain Innovation B.V., Universiteitssingel 40, 6201 BC Maastricht, The Netherlands*

Manuscript information

Title: Neural correlates of conditional and unconditional trust in two-person reciprocal exchange

Authors: Frank Krueger, Kevin McCabe, Jorge Moll, Nikolaus Kriegeskorte, Roland Zahn, Maren Strenziok, Armin Heinecke, Jordan Grafman

Corresponding author

Jordan Grafman, Ph.D.

Cognitive Neuroscience Section

National Institute of Neurological Disorders and Stroke

Building 10, Room 7D43, MSC 1440

National Institutes of Health

Bethesda, Maryland 20892-1440, USA

Phone: 301-496-0220

Fax: 301-480-2909

E-mail: GrafmanJ@ninds.nih.gov

Keywords

neuroeconomics, trust game, social attachment, fMRI, septal area, ventral tegmental area

Number of words, figures and tables

Main text: 2242 words

Methods: 297 words

Figure legends: 710 words

Summary paragraph: 158 words

Number of figures/ tables: 4/0 (manuscript); 6/2 (Supplementary Information)

Trust is an essential feature of human social life. However, the underlying brain mechanisms of conditional and unconditional trust in social reciprocal exchange are still obscure. Here we use hyper-functional magnetic resonance imaging, in which two strangers interacted online with one another in a sequential reciprocal trust game while their brains were simultaneously scanned. We show that the paracingulate cortex is critically involved in building a trust relationship by inferring another person's intentions. This more recently evolved brain region can be differently engaged to interact with more primitive neural systems in maintaining conditional and unconditional trust in a partnership. Conditional trust selectively activated the ventral tegmental area, a region linked to the evaluation of expected and realized reward, whereas unconditional trust selectively activated the septal area, a region linked to social attachment behavior. The interplay of these neural systems supports reciprocal exchange that operates beyond the immediate spheres of kinship, one of the distinguishing features of the human species.

Unlike other species, humans are trustful and cooperate with genetically unrelated strangers, with individuals they will never meet again, or even when reputation and gains are absent[1,2]. Recent studies in experimental economics and social neuroscience have started to explore the neurobiology of trust [2-6] and cooperation [7-10] in reciprocal exchange. Reciprocal behaviour allows the formation of partnerships that can produce mutual advantages for cooperators and thus can be selected for maximizing evolutionary fitness[11]. Reciprocity generally involves a first mover who must trust another person in order to give the other person an opportunity to reciprocate[12]. Typically in a partnership, the person who moves first will vary frequently. In laboratory experiments, trusting behaviour can be reliably reproduced[13,14] although with significant individual variation with respect to both experience[3,15] and context[5,16].

In this paper we look at first movers' decisions to trust. Trusting is always risky given the unpredictability of the intentions of the partner in a social exchange[17]. A trust relationship is built on each partner's decisions to trust and reciprocate. To build a trust relationship, partners must learn that they can depend on each other. One model of this process is the goodwill accounting model[18], which is based on the empirical practice of taking into account the value of ongoing partnerships. Partners accumulate goodwill towards each other and evaluate this against the constantly changing risk of defection. Without balanced goodwill, partners cannot synchronize their mutual cooperation. In this regard, individuals can use one of two strategies that imply different benefits and costs[19]: (i) conditional trust or (ii) unconditional trust.

Conditional trust assumes that one's partner is self-interested and estimates the expected value of one's strategy with respect to the benefits of cooperating, the risk of defection, and the future value of past decisions; it causes less balanced goodwill and results in greater variance in cooperative decisions and, therefore, is cognitively more costly to maintain. In contrast, unconditional trust assumes that one's partner is trustworthy and updates the value of one's partner with respect to their characteristics and past performance; balanced goodwill occurs more quickly allowing the partners to attain high levels of synchronicity in their decisions and, therefore, is cognitively less costly to maintain. In this paper, an examination of functional brain activity supports the hypothesis that the preferential activation of different neuronal systems implements these two trust strategies.

We employed event-related hyper-functional magnetic resonance imaging[20] (MRI), in which two strangers of the same gender (44 participants, 11 female and 11 male pairs) - each in a separate MRI scanner - interacted with one another in a sequential reciprocal trust game while their brains were simultaneously scanned (see Methods, Supplementary Fig. 1).

Participants were asked to make sequential decisions for monetary payoffs (low, medium, or high in cents) presented in a binary game tree (Fig. 1A). The first mover can either quit the game by not trusting the second mover, resulting in a small equal payoff for both; or the first mover can continue the game by trusting the second mover, hoping to receive a better payoff. The second mover can reciprocate the first mover's trust, giving them both a higher payoff, or defect on the first mover's trust, resulting in an even larger payoff for the second mover and a payoff of zero for the first mover. Partners played 36 voluntary trust games and 16 control games. Six blocks of voluntary trust games (6 games per block) were intermixed with four blocks of control games (4 games per block) (Supplementary Fig. 2). In the control games, partners followed the same timeline as in trust games, but they did not have to interact with one another and merely had to choose between lower and higher monetary rewards (Supplementary Fig. 3).

[Insert Fig. 1 about here]

Previous studies used anonymous single or multi-round interactions, in which individuals maintained their roles as first and second mover throughout reciprocal exchange[3,4,8,21]. In their natural environment, however, partners are not anonymous and often alternate in their roles while interacting over long time periods. To improve the ecological validity of the task, we let pairs of strangers play multi-rounds of non-anonymous voluntary trust games while alternating their roles as first and second mover[13] (Fig. 1B). Therefore, trust becomes bidirectional for both partners allowing us to explore partnership building and maintenance while partners develop mental models of one another[6,14]. Previous research has shown that the striatum (caudate head) of second movers in a social reciprocal exchange encodes a signal expecting to see trusting behaviour by their partners[3]. However, the underlying mechanisms of conditional and

unconditional trust in developing a trust partnership are still obscure. The design of our experiment allowed us to address two questions: (i) which brain regions modulate decisions to trust in a partnership and (ii) which brain regions modulate different trust strategies over time in a partnership.

Brain activations and decisions to trust. Data on decisions in voluntary trust games showed that first movers decided to trust significantly more often than not to trust (86% vs. 14%) and second movers reciprocated more often than they defected (77% vs. 8%) (Supplementary Fig. 4A, Supplementary Results). Using a general linear model (GLM) analysis on the multi-subject level, we first sought brain regions whose blood oxygenation level-dependent (BOLD) responses were recruited for decisions to trust. Decisions to trust contrasted with the control condition activated the paracingulate cortex (Fig. 2A, Supplementary table 1). Previous research has shown that the paracingulate cortex not only represents our own thoughts, feelings, and beliefs, but also represents the mental states of other people[6,21-24]. Mentalizing[25] is a unique human characteristic and can be observed only in a most rudimentary form in great apes[26] and has never been observed in monkeys[27]. In building mutual goodwill, partners must infer each other's intentions to determine whether to trust their partners, and whether their partners will reciprocate their trust in the future.

[Insert Fig. 2 about here]

Decisions to trust contrasted with the control condition also activated the septal area (together with the adjoining hypothalamus) (Fig. 2B, Supplementary table 1), a limbic region that has been demonstrated to modulate various aspects of social behaviour including social memory and learning[28]. In addition, the septal area plays a putative role in controlling anterior hypothalamic functions and the release of the neuropeptides vasopressin and oxytocin and itself

contains receptors for those neuropeptides[29-31]. Besides the well-known physiological functions of oxytocin in milk letdown and during labor, oxytocin is a key mediator in facilitating various complex social behaviours, including maternal care[31], pair bonding[32], social recognition[33], and the ability to form social attachment[34-36]. There is evidence that greater first mover trust can be induced in strangers by the nasal administration of synthetic oxytocin[37]. Since synthetic oxytocin increases trust, we surmised that partners recruited the septal area to encode goodwill to maintain their trust partnership. Results from pre- and post-questionnaire ratings support our view demonstrating that partners felt significantly closer to each other and ranked themselves as being more of a partner to the other person after the experiment (Supplementary Fig. 4B).

Trust strategy development. After identifying two distinct regions that underlie decisions to trust in a partnership, we next explored the dynamic role of these regions in supporting conditional and unconditional trust strategies. We arbitrarily divided the experiment into two stages under the assumption that ongoing participation in games during stage I represents partnership building and during stage II, partnership maintenance (Supplementary Fig. 2). In addition, we identified two equal-sized groups based on their decision patterns throughout the experiment: a non-defector group (11 pairs, 6 female pairs) in which neither player ever defected on their partners' decision to trust, and a defector group (11 pairs, 5 female pairs) in which partners experienced some defections during the experiment.

We hypothesized that the non-defector and defector groups would adapt different trust strategies across stages of the experiment. Using a region of interest approach, we derived the parameter estimates from the previous identified paracingulate cortex region to investigate how first movers in the non-defector and defector groups engaged the mentalizing system to build different trust strategies across stages. Further, partners have to balance their goodwill in their

roles as first and second movers to maintain a trust partnership. Using a GLM analysis on the group level, we contrasted decisions to trust with decisions to reciprocate to identify those brain regions that were differently activated for first movers in the non-defector and defector group in maintaining their trust partnership. Finally, we computed brain-to-brain correlations between partners' BOLD amplitude responses in those brain regions when they were first movers in adjacent trials of trust games for the building and maintenance stage (Supplementary Fig. 5). If a correlation reached significance we assumed that partners became "synchronized" in their decision patterns. Results revealed that first movers in the non-defector and defector groups made different use of the mentalizing system resulting in two different neural systems for maintaining unconditional and conditional trust.

Unconditional trust. Unconditional trust assumes that one's partner is trustworthy. During the building stage, first movers in the non-defector group showed higher activation in the paracingulate cortex compared to first movers in the defector group (Fig. 2C). Through mentalizing, partners of this group verified their prior trustworthy assumption, updated the value of one's partner's strategy with respect to their past performance, and maintained a balanced goodwill towards each other allowing them to avoid defections. By developing "better" mental models in this early stage, partners in the non-defector group accumulated sufficient mutual goodwill to become socially attached to each other and adopted an unconditional trust strategy.

[Insert Fig. 3 about here]

During the maintenance stage, the non-defector group showed a higher activation in the septal area compared to the defector group. Across groups, pairs who showed the highest trust-reciprocate history in their decisions also showed the highest activation in this region (Fig. 3A, Supplementary table 2). Furthermore, analyses of pre- and post-scan behavioural ratings

confirmed that only non-defector pairs felt significantly closer to each other and ranked themselves as being more of a partner to the other person after the experiment (Fig. 4A). Through early mentalizing, partners in the non-defector group must have balanced goodwill more quickly allowing them to become synchronized in their septal area's BOLD amplitude responses as first movers (Fig. 3C, Supplementary Fig. 6). Synchronization in the septal area led to social attachment associated with a significant decrease in activation in the paracingulate cortex during the maintenance stage. By adopting this cognitively less costly strategy, decision times over the experiment became significantly faster for the non-defector group. Specifically, decision times accelerated by 20 % for first movers and by 10 % for second movers across stages (Fig. 4B).

[Insert Fig. 4 about here]

Conditional trust. Conditional trust assumes that one's partner is self-interested. During the building stage, first movers in the defector group showed less activation in the paracingulate cortex compared to the non-defector group (Fig. 2C). Through less mentalizing in the building stage, partners in this group produced higher errors in the inferences of second movers' goodwill toward them, resulting in less balanced goodwill and, therefore, in less overall trust compared to the non-defector group. More importantly, they started to trust more in the low payoff games and less in the high payoff games (Fig. 4C). This decision pattern implies that defectors were adapting a conditional trust strategy by evaluating the expected value of one's strategy with respect to the risks and benefits of cooperation.

During the maintenance stage, the defector group showed higher activations in the ventral tegmental area compared to the non-defector group, a region linked to the dopaminergic mesolimbic reward system providing a general reinforcement mechanism to encode expected

and realized reward[38-41]. Across groups, pairs who shared the lowest trust-reciprocate history in their decisions also showed the highest activation in this region (Fig. 3B, Supplementary table 2). By adopting a cognitively more costly strategy, partners in the defector group showed a significant increase in activation in the paracingulate cortex over the experiment. Through more mentalizing in this late stage, first movers in the defector group tried to develop more accurate models about the likelihood of their partner's choices so that they can make a more advantageous decision when to trust. The conditional trust strategy paid off less over time as total earnings decreased for the defector group (but increased for the non-defector group) across stages (Supplementary Fig. 4D).

In conclusion, we applied event-related hyper-fMRI to identify the neural correlates of conditional and unconditional trust when paired strangers interacted with one another in a sequential reciprocal trust game. By designing a non-anonymous, alternating multi-round game, trust became bidirectional and partnership building and maintenance were explored. Our findings extend previous knowledge of the neural basis of cooperation in two-person reciprocal exchange and broaden our understanding of how trust relationships are built and maintained over time. First, the paracingulate cortex is critically involved in building a trust relationship by inferring another person's intentions to predict subsequent behaviour. Second, this more recently evolved brain region can be differently engaged to recruit more primitive neural systems in maintaining conditional and unconditional trust in a partnership. Conditional trust selectively activated the ventral tegmental area, a region linked to the evaluation of expected and realized reward, whereas unconditional trust selectively activated the septal area, a region linked to social attachment behaviour. The interplay of these neural systems supports reciprocal exchange that

operates beyond the immediate spheres of kinship, one of the distinguishing features of the human species.

Methods

Subjects. Forty-four normal volunteers (22 women; age = 28.3 ± 7.1 y/o, education = 17.3 ± 2.2 y/o; mean/SD) took part in the fMRI experiment. All participants were right-handed and native English speakers. Informed consent was obtained according to procedures approved by the NINDS Institutional Review Board.

Data acquisition and analysis. Two 3 Tesla GE MRI scanners equipped with standard circularly polarized head coils were used to acquire single-shot T2* - weighted echoplanar images with BOLD contrast (voxel size = 3.75 x 3.75 x 6 mm) and high resolution T1-weighted structural images. Image analyses were performed using BrainVoyager (Brain Innovation, Maastricht) and custom-written scripts in MATLAB (The MathWorks). Pre-processing steps included: slice-timing and head movement correction, linear trend removal, temporal high-pass filtering, and spatial smoothing (FWHM = 8 mm). General linear models corrected for first-order serial correlation were applied[42] and included regressors created on of participants' decisions in the trust and control games (see Supplementary Methods). Regressor time courses were adjusted for the hemodynamic response delay by convolution with a double-gamma hemodynamic response function[43] and multiple regression analyses were performed to compute parameter estimates. Linear contrasts were applied to the parameter estimates to generate contrast images. Results were derived from random effect analyses by performing t tests on the first level contrast images on the multi-subject and group-level. A priori regions of interest were paracingulate cortex, septal area, dorsal and ventral striatum, and ventral tegmental area. Regions hypothesized to be active were tested for activity using a small volume correction of a sphere of 10-mm radius for

false discovery rate (FDR)[44] with a threshold of $q(\text{FDR}) < 0.05$ (corrected). Non a priori effects were reported using $q(\text{FDR}) < 0.05$ (whole brain analysis). Statistical images were superimposed on a template structural brain in Talairach space[45] and thresholded at $P < 0.005$, uncorrected, with extent threshold of 10 voxels ($t = 3.00$, random effects).

References

1. Smith, V. The two faces of Adam Smith. *Southern Economic Journal*, 1-19 (1998).
2. de Quervain, D.J. et al. The neural basis of altruistic punishment. *Science* **305**, 1254-8 (2004).
3. King-Casas, B. et al. Getting to know you: reputation and trust in a two-person economic exchange. *Science* **308**, 78-83 (2005).
4. Tomlin, D. et al. Agent-specific responses in the cingulate cortex during economic exchanges. *Science* **312**, 1047-50 (2006).
5. Delgado, M.R., Frank, R.H. & Phelps, E.A. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat Neurosci* **8**, 1611-8 (2005).
6. McCabe, K., Houser, D., Ryan, L., Smith, V. & Trouard, T. A functional imaging study of cooperation in two-person reciprocal exchange. *Proc Natl Acad Sci U S A* **98**, 11832-5 (2001).
7. Decety, J., Jackson, P.L., Sommerville, J.A., Chaminade, T. & Meltzoff, A.N. The neural bases of cooperation and competition: an fMRI investigation. *Neuroimage* **23**, 744-51 (2004).
8. Rilling, J. et al. A neural basis for social cooperation. *Neuron* **35**, 395-405 (2002).
9. Singer, T., Kiebel, S.J., Winston, J.S., Dolan, R.J. & Frith, C.D. Brain responses to the acquired moral status of faces. *Neuron* **41**, 653-62 (2004).
10. Singer, T. et al. Empathic neural responses are modulated by the perceived fairness of others. *Nature* **439**, 466-9 (2006).
11. Axelrod, R. & Hamilton, W.D. The evolution of cooperation. *Science* **211**, 1390-6 (1981).
12. Arrow, K.J. *The Limits of Organizations*, (Norton & Company, Inc., New York, 1974).

13. Berg, J., Dickhaut, J. & McCabe, K. Trust, Reciprocity, and Social History. *Games and Economic Behavior* **10**, 122-142 (1995).
14. Camerer, C.F. *Behavioral game theory: Experiments in strategic interactions*, (Princeton University Press, Princeton, NJ, 2003).
15. McCabe, K. & Smith, V. A Two Person Trust Game Played by Naïve and Sophisticated Subjects. *Proc Natl Acad Sci U S A* **97**, 3777-3781 (2000).
16. Burnham, T., McCabe, K. & Smith, V. Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior and Organization* **43**, 57-73 (2000).
17. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785-91 (2003).
18. McCabe, K. & Smith, V. Goodwill accounting in economic exchange. in *Bounded Rationality: The Adaptive Toolbox* (eds. Gigerenzer, G. & Selten, R.) 319-340 (MIT Press, Cambridge, MA, 2001).
19. Williamson, O.E. Calculativeness, Trust and Economic Organization. *Journal of Law and Economics* **36**, 453-486 (1993).
20. Montague, P.R. et al. Hyperscanning: simultaneous fMRI during linked social interactions. *Neuroimage* **16**, 1159-64 (2002).
21. Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E. & Cohen, J.D. The neural correlates of theory of mind within interpersonal interactions. *Neuroimage* **22**, 1694-703 (2004).
22. Amodio, D.M. & Frith, C.D. Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci* **7**, 268-77 (2006).
23. Gallagher, H.L., Jack, A.I., Roepstorff, A. & Frith, C.D. Imaging the intentional stance in a competitive game. *Neuroimage* **16**, 814-21 (2002).

24. Goel, V., Grafman, J., Sadato, N. & Hallett, M. Modeling other minds. *Neuroreport* **6**, 1741-6 (1995).

25. Frith, C.D. & Frith, U. Interacting minds--a biological basis. *Science* **286**, 1692-5 (1999).

26. Povinelli, D.J. & Preuss, T.M. Theory of mind: evolutionary history of a cognitive specialization. *Trends Neurosci* **18**, 418-24 (1995).

27. Cheney, D.L. & Seyfarth, R.M. *How monkeys see the world: inside the mind of another species.*, (University of Chicago Press, Chicago, 1990).

28. Numan, R. (ed.) *The behavioral neuroscience of the septal region*, (Springer, New York, 2000).

29. Loup, F., Tribollet, E., Dubois-Dauphin, M. & Dreifuss, J.J. Localization of high-affinity binding sites for oxytocin and vasopressin in the human brain. An autoradiographic study. *Brain Res* **555**, 220-32 (1991).

30. Powell, E.W. & Rorie, D.K. Septal projections to nuclei functioning in oxytocin release. *Am J Anat* **120**, 605-10 (1967).

31. Insel, T.R. & Young, L.J. The neurobiology of attachment. *Nat Rev Neurosci* **2**, 129-36 (2001).

32. Insel, T.R. & Shapiro, L.E. Oxytocin receptor distribution reflects social organization in monogamous and polygamous voles. *Proc Natl Acad Sci U S A* **89**, 5981-5 (1992).

33. Choleris, E. et al. An estrogen-dependent four-gene micronet regulating social recognition: a study with oxytocin and estrogen receptor-alpha and -beta knockout mice. *Proc Natl Acad Sci U S A* **100**, 6192-7 (2003).

34. Moll, J. et al. Human fronto-mesolimbic networks guide decisions about charitable donation. *Proc Natl Acad Sci U S A* **103**, 15623-8 (2006).

35. Bartels, A. & Zeki, S. The neural correlates of maternal and romantic love. *Neuroimage* **21**, 1155-66 (2004).

36. Aron, A. et al. Reward, motivation, and emotion systems associated with early-stage intense romantic love. *J Neurophysiol* **94**, 327-37 (2005).

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

38. Schultz, W., Dayan, P. & Montague, P.R. A neural substrate of prediction and reward. *Science* **275**, 1593-9 (1997).

39. Schultz, W. Predictive reward signal of dopamine neurons. *J Neurophysiol* **80**, 1-27 (1998).

40. Fiorillo, C.D., Tobler, P.N. & Schultz, W. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898-902 (2003).

41. Schultz, W. & Dickinson, A. Neuronal coding of prediction errors. *Annu Rev Neurosci* **23**, 473-500 (2000).

42. Friston, K.J., Holmes, A.P., Price, C.J., Buchel, C. & Worsley, K.J. Multisubject fMRI studies and conjunction analyses. *Neuroimage* **10**, 385-96 (1999).

43. Friston, K.J. et al. Event-related fMRI: characterizing differential responses. *Neuroimage* **7**, 30-40 (1998).

44. Genovese, C.R., Lazar, N.A. & Nichols, T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* **15**, 870-8 (2002).

45. Talairach, J. & Tournoux, P. *Co-Planar Stereotaxic Atlas of the Human Brain*, (Thieme Medical Publishers, New York, 1988).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors are grateful to E. Wassermann for performing the neurological exams. We thank N. Armstrong and M. Ulrich for their help in various stages of this project. We are grateful to A. Butz for helping programming the fMRI experiment as well as U. Hasson and H. Gelbard-Sagivfor for making available MATLAB scripts. The authors thank J. Bodurka and S. Marrett for technical assistance for the hyper-fMRI setup and E. Condon for scanning assistance. The authors are grateful for V. Smith's and E. Huey's helpful comments. The work was supported in part by a postdoctoral NINDS competitive fellowship award for F.K. and the Intramural Research Program of the CNS/ NINDS/ NIH.

Competing financial Interests. The authors declare that they have no competing financial interests.

Authors Contributions. F.K., K.M., J.M. and J.G. conceived the experiment. F.K., J.M., R.Z. and M.S. carried out data acquisition. F.K., N.K., and A.H. performed the data analysis. F.K., K.M., and J.G. co-wrote the paper.

Correspondence and requests for materials should be addressed to J.G. (grafmanj@ninds.nih.gov).

Figure 1 Experimental design. **a,** Voluntary trust game. Partners made sequential decisions as first mover (M1) and second mover (M2) for payoffs in cents (c: [c_{M1},c_{M2}]) presented in a binary decision tree. M1 can choose left (non-trust) and quit the game with a small payoff for M1 and M2 (e.g., [5,5]) or can choose right (trust) to continue the game. M2 can then choose left (reciprocate) giving them both a higher payoff (e.g., [10,15]) or choose right (defect) resulting in an even larger payoff to M2 and a payoff of zero to M1 (e.g., [0,25]). Payoffs (p_1-p_6) were split into three types: low (p_1 - p_2), medium (p_3 - p_4), and high (p_5 - p_6). **b,** Timeline for a single trust game. Partners were introduced by seeing each other via webcam and digital photographs were taken to be used for game trials. A 2-s introductory screen informed partners of the role that they were playing (M1 or M2). M1 saw the game tree, had to make a decision (non-trust or trust) within 6 s, and waited 6s for M2's decision while seeing a blank screen. M2 saw a blank screen for 6 s, saw the game tree with M1's decision, and had to make a decision (reciprocate or defect) within 6 s. If M1 had chosen not to trust M2, the game was over and M2 saw M1's decision for 6 s. Partners saw the outcome of the game for 4 s followed by a blank screen with a jittered inter-stimulus interval of 2 s to 6 s.

Figure 2 Brain responses for decisions to trust. **a,** Trust building. Decisions to trust contrasted with the control condition activated the paracingulate cortex (Brodmann's areas, BA 9/32; peak voxel: 5,39,22). **b,** Trust maintenance. Decisions to trust contrasted with the control condition activated the septal area (peak voxel: -4,4,-3). **c,** Trust development. First movers in the non-defector and defector groups made different use of the mentalizing system across stages ($F(1,42) = 9.14$, $P < 0.004$). The non-defector group showed a higher activation (parameter estimates, ± s.e.m.) in the paracingulate cortex compared to the defector group in the building stage ($t(42) = 2.72$, $P < 0.010$). The non-defector group showed a decrease in activation

($t(21) = 2.10$, $P < 0.048$), while the defector group showed an increase in activation ($t(21) = -2.18$, $P < 0.041$) in the paracingulate cortex across stages.

Figure 3 Brain responses for trust maintenance. **a,** Unconditional trust. In the non-defector group, decisions to trust contrasted with decisions to reciprocate revealed a higher activation in the septal area (peak voxel: 1,2,-4) compared to the defector group. Pairs who showed the highest trust-reciprocate history (frequency) in their decisions also showed the highest activation (parameter estimates) in the septal area ($r = 0.59$, $P < 0.004$). **b,** Conditional trust. In the defector group, decisions to trust contrasted with decisions to reciprocate revealed a higher activation in the ventral tegmental area (peak voxel: 2,-20,-13) compared to the non-defector group. Pairs who showed the lowest trust-reciprocate history (frequency) in their decisions also showed the highest activation (parameter estimates) in the ventral tegmental area ($r = -0.63$, $P < 0.002$). **c,** Brain-to-brain correlation (± s.e.m.). In the non-defector group, brain-to-brain correlations increased in the septal area across stages ($t(10) = -2.40$, $P < 0.038$). In the maintenance stage, partners in the non-defector group became synchronized in their septal area's BOLD amplitudes as first movers in adjacent trials of trust games ($r = 0.27$, $P < 0.005$).

Figure 4 Behavioural results for trust development. **a,** Pre- and post-experiment ratings (± s.e.m.). Before and after scanning, partners were asked to rate their closeness and partnership to one another on Likert-scales. Partners in the non-defector group felt closer to each other ($t(21) = -3.24$, $P < 0.004$) and ranked themselves as more of a partner to the other person ($t(21) = -2.99$, $P < 0.007$) after the experiment. **b,** Decision times (± s.e.m.). Decision times for trust games became faster for the non-defector group across stages ($F(1,21) = 5.86$, $P < 0.025$). Decision times accelerated by 20 % for first movers ($t(21) = 5.15$, $P < 0.001$) and by 10 % for second movers ($t(21) = 2.71$, $P < 0.013$). **c,** Behavioural choices (± s.e.m.). Trust in the non-

defector group was higher than in the defector group ($F(1,42) = 26.62$, $P < 0.001$) and increased across stages ($F(1,21) = 5.86$, $P < 0.025$). Trust in the defector group decreased across stages ($F(1,21) = 4.37$, $P < 0.048$) and depended on the payoff type ($F(2,42) = 9.57$, $P < 0.001$). In the maintenance stage, trust in this group occurred more often in the low payoff games compared to the medium and high payoff games ($F(1,21) = 23.25$, $P < 0.001$) and in the medium compared to the high payoff games ($F(1,21) = 4.91$, $P < 0.038$).