King-Casas, B; Sharp, C; Lomax-Bream, L; Lohrenz, T; Fonagy, P; Montague, PR; (2008) The rupture and repair of cooperation in borderline personality disorder. **Science**, 321 (5890) 806 - 810. 10.1126/science.1156902.

ARTICLE

Computational substrates of a social exchange disease: the rupture and repair of cooperation in borderline personality disorder

Brooks King-Casas[1,2], Carla Sharp[2], Laura Lomax[2],

Terry Lohrenz[1], Peter Fonagy[2,3], P. Read Montague[1,2]

[1]*Computational Psychiatry Unit &*

Department of Neuroscience

[2]*Menninger Department of Psychiatry and Behavioral Sciences*

Baylor College of Medicine

1 Baylor Plaza, Houston, TX 77030

[3]*Department of Psychology*

University College London

Gower Street, London, UK

WC1E 6BT

ABSTRACT

To sustain or repair cooperation, adaptive social creatures must understand the responses expected from social gestures and the consequences resulting when such norms are violated by accident or intent. We have applied this normative perspective to participants with a diagnosis of borderline personality disorder (BPD) playing a mathematically portrayed multi-round exchange game with a healthy partner. Subjects with BPD showed a profound incapacity to maintain cooperation, and they were impaired in their capacity to repair it based on a quantitative measure of coaxing. In a separate experiment, an adaptive computer agent, designed to play the game like a BPD subject, provided strong evidence that BPD subjects cause the break in cooperation. Healthy subjects showed a strong linear relationship between

their partner's offer level and the response of the anterior insular cortex – a region known to respond to norm violations across a number of dimensions. In contrast, BPD subjects' anterior insula did not differentiate offer levels despite displaying the same average response. These neural and behavioral data suggest that social exchange norms in BPD subjects are pathologically perturbed or missing altogether. This game-theoretic approach to psychopathology may open doors to new ways of characterizing and understanding a range of mental illnesses.

Theoretical and empirical work on cooperation has made tremendous strides by using game theory to provide a normative mathematical setting for understanding social exchange *(1-7)*. This same game-theoretic framework has exposed some of the basic computations underlying social interaction and identified neural correlates of cooperation, reciprocity, and social signaling *(8-22)*. Collectively, this work raises the possibility of a new approach to characterizing and understanding psychopathology from a normative perspective. In conditions ranging from psychosis to developmental and personality disorders, afflicted individuals often display a dramatically perturbed capacity to model others and to sense and respond appropriately to the social signals they emit *(23-25)*. Consequently, using normative game-theoretic probes of social signaling in identified psychopathologies offers the opportunity to understand some of the components of these disorders in terms of malfunctioning computations *(26, 27)*.

In this paper, we selected subjects based on their capacity to maintain stable interpersonal relationships and used a multi-round economic exchange game (**Fig. 1A**) to probe two fundamental components underlying the capacity to sustain a successful social exchange: cooperation and its repair. Specifically, we studied a group of control individuals and a second group of individuals diagnosed with borderline personality disorder (BPD), a psychiatric disorder characterized by unstable relationships, affective dysregulation, and a general incapacity to trust appropriately the actions and (possible) motives of others *(28, 29)*

Successful cooperation between two agents requires a range of intact computations including the capacity to sense and to value social gestures exchanged with one's partner *(30-32)*, as well as respond with strategies that promote individual and/or shared goals. However, such cooperative exchange is fragile, and can easily be ruptured whether by accident (e.g., through impoverished models of social partners) or intent *(33)*. Successful repair of broken cooperation in an iterated exchange requires the capacity to coax one's partner back into cooperation

through the medium of generous gestures *(33-39)*. In an economic exchange game, such gestures are encoded as money units thus exposing their immediate cost to each subject. However, for such gestures to be meaningful, cooperating individuals must possess *correct* fairness norms for what is expected from them and their partners, and they must sense when these norms have been significantly violated *(40, 41)*. In this paper, we pursued the hypothesis that social exchange norms of subjects with borderline personality disorder are either pathologically perturbed or missing altogether, thus preventing sustained cooperation or its repair after it breaks down.

Before scanning, all participants (healthy subject group and BPD group) underwent diagnostic and symptom assessment, and were matched on demographic variables including sex, age, education and verbal IQ (**see Table S1 and SOM**). Within each trust exchange, cooperation occurs when an investor and trustee act in a manner that mutually benefits both players. For example, if an investor sends $20 to a trustee and the trustee splits the tripled investment ($60) with the investor, both the investor and trustee profit. The investor earns $10 more, and the trustee earns $30 more, than if the investor had sent nothing. However, if a trustee does not repay at least the amount invested, the investor accrues no benefit from the exchange, likely triggering smaller subsequent investments. Thus, increased cooperation is seen with increased money exchanged across the course of the 10-round game. In the current study, subjects diagnosed with BPD played the trustee role against a healthy investor; as expected, each game began normally with the healthy investor showing a level of trust comparable to that of healthy investors playing healthy trustees (**Fig. 1A and SOM**).

Individuals with BPD send signals that break cooperation. In early rounds of this multi-round game, investment levels did not differ between subjects partnered with BPD trustees and subjects partnered with healthy trustees. However, in late rounds of the game, investments

were significantly lower for dyads with a BPD trustee relative to control dyads with a healthy trustee (**Fig. 1B**). This definitive downward shift in investment levels for dyads with BPD trustees reflects a break in cooperation (**see fig. S2 for round-by-round summary of the same data and footnote 1 for regression analysis of dyadic interaction**). We strongly suspected that this breakdown was caused by the social signals emitted by the BPD trustee and were not due to a sampling of healthy investors that happened to be uncooperative. To test this claim empirically, we developed two adaptive computer agents - one that played the multi-round trust game like a BPD trustee and one that played like a healthy trustee (method illustrated in **Fig. 1B**). In a separate within-subject experimental design, healthy investors played each computer agent (n=68; **Fig. 1C**). To account for learning effects, this large group of investors was split into two groups and counterbalanced over which computer agent (healthy or BPD) the human investor played first (**SOM**). As shown in **Fig. 1C**, the healthy computer trustee elicited investment levels consistent with data collected from two healthy interacting humans in **Fig. 1A**. However, the BPD-like computer agent elicited significantly lower levels of cooperation in the same investors. This separate set of experiments strongly suggests that the rupture in cooperation shown in **Fig. 1A** is due to the signals emitted by the BPD subjects rather than being an artifact contingent on some other unidentified variable.

Coaxing pays and BPD subjects do it much less frequently. Although the BPD subjects showed a consistent tendency to rupture cooperation with their healthy partners, the issue of broken cooperation arises during all interpersonal exchange. Individuals bring to any two-party interaction variability in their fairness norms, variable sensitivities to deviations from these norms, and variable responses to these deviations. These issues highlight a feature that plagues all social exchange – breakdowns in cooperation are common, and successful social agents must sense (or even anticipate) these failures and select actions to repair them *(33-39)*. Consequently, we specifically sought to identify strategic responses to cooperation breakdowns

that allow healthy trustees to maintain high levels of cooperation into the later rounds of the exchange. We focused on rounds in which cooperation was low, that is, rounds in which investments of $5 or less were made. Low investments act as a viable proxy for broken cooperation because in this economic exchange game and other related games *(1, 10, 13, 15, 16, 42)*, opening investments to a responder (here the trustee) are typically large. For investments to become small, a responder has to 'prove' that they are not worthy of a large offer by acting in an untrustworthy manner. However, such circumstances offer the trustee an opportunity to repair the broken cooperation. If trustees repay a large fraction of the tripled investment, they signal their trustworthiness in the presence of low offers and may garner greater investments on subsequent rounds (**Fig. 2A**). We term this behavior 'coaxing' and show in **Fig. 2B** that coaxing pays dividends for 4 rounds into the future. In this same figure, we summarize the results for low offers when the trustees do not coax – there is no future payback. So in the context of this multi-round game, coaxing repairs cooperation and pays off generously for the coaxer. Remarkably, comparison of healthy trustees to BPD trustees found healthy players to be almost twice as likely as BPD players to coax in the presence of low offers (i.e., repay a third or more of the investment; **Fig. 2C**).

BPD subjects have perturbed norm deviation responses in their insular cortex. These behavioral data with humans and computer agents suggest strongly that BPD subjects emit responses that cause cooperation to rupture, and, once ruptured, they show substantially diminished rates of coaxing (generous gestures) to repair the break (**Fig. 2C**). A question naturally arises about whether there is a comprehensible and consistent neural correlate of the rupture in cooperation that serves to cue the repair (or not) of cooperation. To investigate the possible neural origins of failures in cooperation and the lack of coaxing, we first sought neural responses in healthy trustee brains that were stronger in response to the revelation of small investments than to the revelation of large investments. Identified regions included bilateral

anterior insula, medial frontal gyrus, bilateral inferior parietal lobule (BA 40), and left inferior frontal gyrus (**see table S3**). In contrast to these findings in healthy subjects, *no regions* were significantly related to investment size among individuals with borderline personality disorder. In earlier work using this same multi-round trust game *(13)* and single-shot ultimatum games *(19)*, small offers increase the probability of defection in the partner and such behavior is preceded by increased activity in the anterior insular cortex (also see *(43)* for review). Consequently, we pursued a region of interest analysis of the anterior insular cortex (**Fig. 3**). A dramatic difference emerged between healthy trustees and BPD subjects. In healthy subjects, there was a strong linear relationship between the size of the offer sent by the investor and the response of the anterior insular cortex. In BPD subjects, the insular cortex was activated at the same level across all investment levels, suggesting that their brains either perceive all offers as threatening (like a low offer) or that they simply cannot differentiate the intended meaning of different offer levels (see discussion).

Discussion

Using a ten-round iterated exchange of trust and event-related fMRI, we tracked breakdowns in cooperation in a large cohort of 78 *pairs* of subjects, composed of one group of healthy social agents (n = 36 pairs) and another group of individuals with borderline personality disorder (n = 42 pairs), a psychiatric illness characterized by an inability to maintain stable interpersonal relationships. In healthy subject pairs, we find that when cooperation fails, norm-violation responses in the anterior insula are observed (**Fig. 3**), and cooperative exchange is restored through a behavioral mechanism of cooperation repair that we here labeled coaxing (**Fig. 2A**). In contrast to healthy trustees, we find that BPD subjects emit signals that result in a rupture in cooperation (**Fig. 1C**), and they display a significantly diminished cooperation-repair response of coaxing (half the rate of healthy subjects (**Fig. 2C**). The mitigated coaxing response in BPD

subjects is significant because such coaxing behavior effectively reestablishes cooperation (**Fig. 2B**). Furthermore, coaxing in the context of this game is consistent with the generous or forgiving strategies identified by evolutionary simulations of cooperative exchange, as well as previous empirical results of iterated prisoner's dilemma games *(33-39)*.

Using two adaptive computer agents and a separate large cohort of healthy investors (n = 68), we showed that the behavioral signal distributions emitted by BPD subjects were sufficient on their own to disrupt cooperation while a similar computer agent playing like a healthy trustee maintained cooperation with its partner (**Fig. 1C**). But the most telling neural feature observed in the BPD group was the lack of an insular response that differentiated offer levels, yet displayed the same average activation level as healthy trustees across all offer levels (**Fig. 3**). While the insular cortex has traditionally been associated with pain perception and representation *(44, 45)*, responses in the anterior insula have recently been identified in social interactions where norms are violated *(19, 21, 46, 47)*. In an ultimatum game, in which one player offers to split an endowment with a second player, small offers are perceived as unfair and typically refused *(1, 48, 49)*. In such instances, anterior insula activity both scales negatively with offer size and predicts whether the offer is subsequently rejected *(19)*. In a related finding, anterior insula activity of *observers* is greater when a punishment is applied to players perceived as fair relative to players perceived as unfair *(20)*. In non-social tasks, activity in the anterior insula has been found to encode representations of risk and uncertainty about decision outcomes *(50-56)*. The association of the insula with a representation of outcome variance suggests that the insula may encode the distribution of likely outcomes in social interactions, that is, responses in the anterior insula may indicate social norm violations within interpersonal contexts. Furthermore, violations of such norms could serve as control signals that update expectations about social partners or at least inform learning about a subset of parameters associated with one's partner. This possibility is supported by the work of Preuschoff and Bossaerts who have recently reported

prediction errors of risk to evoke strong responses in the bilateral anterior insula, consistent with this speculation *(55-57)*.

Taken together, these data support the hypothesis that the strong negative relationship between offer size and activity in the anterior insula seen among healthy trustees reflects the perceived violation of social norms in the two-party trust exchange. The apparent insensitivity of the BPD subjects' insula to offer level size suggests two possibilities along these lines: 1) monetary reward is not reinforcing to individuals with BPD; or, 2) low offers are not perceived to be a violation of social norms to individuals with BPD. Previous work has shown monetary reward to be a potent reinforcer in natural and laboratory settings in this group *(58)*, making the former less likely. Furthermore, studies of interpersonal and emotional processing in BPD suggest that this group holds negative expectations of social partners and exhibits negative evaluative biases *(59, 60)*, consistent with the presence of atypical social norms.

To compare social norms of healthy trustees to BPD trustees within the current study, a self-report measure of trust *(61)* was used to gauge expectations across a variety of social situations and with a variety of social agents. Individuals with BPD expressed significantly lower levels of self-reported trust relative to healthy controls ($p < .001$; **Fig. 4**), a finding that agrees with the diminished trust exhibited by this group on the economic exchange. Together, these results suggest that the diminished insula response to low offers does indeed reflect atypical social norms in this group. Put another way, the low offers from social partners do not violate the social expectations of individuals with BPD, accounting for the diminished insula response in the BPD group.

While a number of studies have utilized resting-state and emotional challenge paradigms to investigate personality disorders, the current study represents the first large-scale investigation

of interpersonal behavior among individuals diagnosed with an Axis-II psychiatric disorder using a normative exchange game. The critical role that interpersonal deficits play across a variety of such disorders, along with the substantive contribution that neuroimaging studies have made in elucidating the etiology of Axis-I disorders, recommend future imaging studies of active social exchanges between individuals with and without social pathologies *(26, 27, 62)*.

References

Footnote X. To explore whether lower levels of tit-for-tat reciprocity may have contributed to such failures of cooperation, linear regressions of behavior between partners were carried out. Consistent with previous work, reciprocity was found to be a strong predictor of subsequent increase or decreases in amount of money sent (**table S2**). Reciprocity is defined as the fractional change in money sent across rounds by one player in response to the fractional change in money sent across rounds by their partner. Thus, investor reciprocity on round j was calculated as $\Delta I_j - \Delta R_{j-1}$, where ΔI_j indicates the fractional change in investment from round $j - 1$ to round j and ΔR_{j-1} indicates the fractional change in repayment from round $j - 2$ to round $j - 1$. While deviation in perfect tit-for-tat behavior was a significant predictor of change in partner behavior across groups, the strength of this relationship did not differ between pairs that included healthy trustees and pairs that included BPD trustees (**table S2**), suggesting sensitivity and responsivity to tit-for-tat behavior is intact among individuals with BPD.

1. C. Camerer, *Behavioral Game Theory* (2003).

2. R. Axelrod, W. D. Hamilton, *Science* **211**, 1390 (1981).

3. H. Gintis, *Game Theory Evolving* (2000).

4. M. A. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life* (2006).

5. R. L. Trivers, *Q. Rev. Biol.* **46**, 35 (1971).

6. J. Maynard-Smith, *Cambridge University Press. Cambridge* (1982).

7. J. von Neumann, O. Morgenstern, *2nd Ed.* (1947).

8. D. J. -. De Quervain *et al.*, *Science* **305**, 1254 (2004).

9. J. Decety, P. L. Jackson, J. A. Sommerville, T. Chaminade, A. N. Meltzoff, *NeuroImage* **23**, 744 (2004).

10. M. R. Delgado, R. H. Frank, E. A. Phelps, *Nature Neuroscience* **8**, 1611 (2005).

11. N. I. Eisenberger, M. D. Lieberman, K. D. Williams, *Science* **302**, 290 (2003).

12. W. T. Harbaugh, U. Mayr, D. R. Burghart, *Science* **316**, 1622 (2007).

13. B. King-Casas *et al.*, *Science* **308**, 78 (2005).

14. D. Knoch, A. Pascual-Leone, K. Meyer, V. Treyer, E. Fehr, *Science* **314**, 829 (2006).

15. K. McCabe, D. Houser, L. Ryan, V. Smith, T. Trouard, *Proceedings of the National Academy of Sciences of the United States of America* **98**, 11832 (2001).

16. F. Krueger *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20084 (2007).

17. J. K. Rilling *et al.*, *Neuron* **35**, 395 (2002).

18. J. K. Rilling, A. G. Sanfey, J. A. Aronson, L. E. Nystrom, J. D. Cohen, *NeuroReport* **15**, 2539 (2004).

19. A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, J. D. Cohen, *Science* **300**, 1755 (2003).

20. T. Singer *et al.*, *Nature* **439**, 466 (2006).

21. M. Spitzer, U. Fischbacher, B. Herrnberger, G. Grön, E. Fehr, *Neuron* **56**, 185 (2007).

22. D. Tankersley, C. J. Stowe, S. A. Huettel, *Nature Neuroscience* **10**, 150 (2007).

23. M. Sprong, P. Schothorst, E. Vos, J. Hox, H. Van Engeland, *British Journal of Psychiatry* **191**, 5 (2007).

24. M. Brüne, U. Brüne-Cohrs, *Neuroscience and Biobehavioral Reviews* **30**, 437 (2006).

25. J. Rogers, E. Viding, R. J. Blair, U. Frith, F. Happé, *Psychological Medicine* **36**, 1789 (2006).

26. P. H. Chiu *et al.*, *Neuron* **57**, 463 (2008).

27. J. K. Rilling *et al.*, *Biological Psychiatry* **61**, 1260 (2007).

28. K. Lieb, M. C. Zanarini, C. Schmahl, M. M. Linehan, M. Bohus, *Lancet* **364**, 453 (2004).

29. American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR (American Psychiatric Association, Washington, DC, ed. 4, text revised, 2000).

30. C. D. Frith, U. Frith, *Brain Research* **1079**, 36 (2006).

31. T. Singer, *Neuroscience and Biobehavioral Reviews* **30**, 855 (2006).

32. A. G. Sanfey, G. Loewenstein, S. M. McClure, J. D. Cohen, *Trends Cogn. Sci.* **10**, 108 (2006).

33. R. Axelrod, D. Dion, *Science* **242**, 1385 (1988).

34. J. Bendor, R. M. Kramer, S. Stout, *The Journal of Conflict Resolution* **35**, 691 (1991).

35. D. Fundenberg, E. Maskin, *Am. Econ. Rev.* **80**, 274 (1990).

36. J. Wu, R. Axelrod, *Journal of Conflict Resolution*, 183 (1995).

37. M. A. Nowak, K. Sigmund, *Nature* **355**, 250 (1992).

38. M. A. Nowak, *Science* **314**, 1560 (2006).

39. C. Wedekind, M. Milinski, *Proceedings of the National Academy of Sciences of the United States of America* **93**, 2686 (1996).

40. E. Fehr, C. F. Camerer, *Trends Cogn. Sci.* (2007).

41. C. F. Camerer, E. Fehr, *Science* **311**, 47 (2006).

42. J. Henrich *et al.*, *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from 15 Small-Scale Societies* (2004).

43. B. Seymour, T. Singer, R. Dolan, *Nature Reviews Neuroscience* **8**, 300 (2007).

44. A. D. Craig, *Current Opinion in Neurobiology* **13**, 500 (2003).

45. A. D. Craig, *Annual Review of Neuroscience* **26**, 1 (2003).

46. P. R. Montague, T. Lohrenz, *Neuron* **56**, 14 (2007).

47. M. P. Paulus, *Science* **318**, 602 (2007).

48. E. Fehr, K. M. Schmidt, *Quarterly Journal of Economics* **114**, 817 (1999).

49. W. Güth, R. Schmittberger, B. Schwarze, *Journal of Economic Behavior and Organization* **3**, 367 (1982).

50. G. S. Berns *et al.*, *Science* **312**, 754 (2006).

51. H. D. Critchley, C. J. Mathias, R. J. Dolan, *Neuron* **29**, 537 (2001).

52. S. A. Huettel, C. J. Stowe, E. M. Gordon, B. T. Warner, M. L. Platt, *Neuron* **49**, 765 (2006).

53. A. Simmons, S. C. Matthews, M. P. Paulus, M. B. Stein, *Neuroscience Letters* **430**, 92 (2008).

54. M. P. Paulus, C. Rogalsky, A. Simmons, J. S. Feinstein, M. B. Stein, *NeuroImage* **19**, 1439 (2003).

55. K. Preuschoff, P. Bossaerts, S. R. Quartz, *Neuron* **51**, 381 (2006).

56. K. Preuschoff, P. Bossaerts., *Adding prediction risk to the theory of reward learning* (2007).

57. K. Preuschoff, S. Quartz, P. Bossaerts, *Journal of Neuroscience* (2008).

58. D. M. Dougherty, J. M. Bjork, H. C. G. Huckabee, F. G. Moeller, A. C. Swann, *Psychiatry Research* **85**, 315 (1999).

59. K. M. Putnam, K. R. Silk, *Development and Psychopathology* **17**, 899 (2005).

60. S. Sieswerda, A. Arntz, M. Wolfis, *Journal of Behavior Therapy and Experimental Psychiatry* **36**, 209 (2005).

61. J. B. Rotter, *Journal of Personality* **35**, 651 (1967).

62. A. Todorov, L. T. Harris, S. T. Fiske, *Brain Research* **1079**, 76 (2006).

Figure Legends

Fig. 1. Cooperation fails across rounds when individuals with BPD engage in a repeated exchange of trust.

(A) Cooperation across 10-round trust game. Investors are endowed with $20 at the start of each of 10 rounds. Investors can invest any portion of the endowment with their partner. The invested amount is tripled before being passed to the trustee, who can repay any portion of the tripled investment. Roles are kept constant for all rounds of the game. Trustees accrue earnings during each round by keeping a portion of the tripled investment, while investors accrue earnings both by the portion of the original endowment kept, as well as any repayments made by their partner. For additional details of the experimental design, see **fig. S1**, **table S1** and **SOM**. Among 36 healthy investors (gray) paired with healthy trustees (gray), investments were large and sustained across early (1-5) rounds and late (6-10) rounds of the game. However, among 42 healthy investors (gray) paired with 42 trustees with borderline personality disorder (BPD; red), a decrease in investment level from early to late rounds of the game indicates a failure in cooperation across the iterated exchange. Mean percent invested and SEM are plotted.

(B) Adaptive computer trustee. Decisions of the adaptive computer partners were generated using a k-nearest neighbors sampling algorithm. Specifically, choices of a 'computer trustee' were generated by: (a) identifying interactions within a database of 10-round trust game behavior that are similar (i.e., 7 dyads in the database with smallest Euclidean distance from the vector of 5 previous choices – last 3 investments and last 2 repayments) and then (b) randomly sampling from the distribution of next repayments of those nearest neighbors in order to generate 'computer trustee' choices. In doing so, 'computer trustees' behave as average

healthy human trustees when sampling from the healthy database (n = 36), and behave as average BPDs when sampling from a database of BPD dyads (n = 42).

(C) To confirm that the break in cooperation observed in late rounds among BPD dyads can be accounted for by behavioral signals sent by BPD players, a separate cohort of healthy investors (n = 68) played two adaptive computer trustees in a within-subject design. In one condition, healthy investors played 10 rounds with a 'healthy' computer trustee; in a second condition, the same investor played 10 rounds with a 'BPD' computer trustee. The order of play was counterbalanced, and the effect of 'computer trustee' type on level of cooperation (mean investment ± SEM) is depicted. Consistent with results depicted in **Fig. 1A**, we found BPD 'computer trustee' behavior elicited lower levels of cooperation than healthy 'computer trustee' behavior within the same human investors.

Fig. 2.

(A) Coaxing by trustee engenders trust from investors. Cooperation within the iterated trust game is expressed as high levels of investment combined with equitable sharing of investment returns (see **fig. S2** for normative behavior). When cooperation falters (i.e., when investments are low), trustees can rebuild trust and cooperation by 'coaxing' back higher and higher investment levels. In this schematic representation, the investor entrusts only a small portion of their endowment ($5 of $20 points) to their partner. The $15 received by the trustee is far less than the $60 that would have been received if the investor had sent the entire endowment. Thus, in order to maximize their gains, a trustee must induce cooperation in a repeated exchange. The trustee can coax back trust from their partner by repaying a large proportion of the tripled investment -- signaling their own trustworthiness and thus eliciting larger subsequent investments. Conversely, if the trustee does not coax and instead keeps a large proportion of

the tripled investment, the investor is likely to invest less on subsequent rounds and cooperation will continue to devolve.

(B) Coaxing pays off. The effect of such coaxing is summarized in the effect of repayment on investments in subsequent rounds. This analysis is restricted to low investments (all rounds with investments <$5). Gray bars depict mean ± SE investment level following small repayments (repayment less than 1/3 of the tripled investment; 'no coax' condition), while red bars depict investment level following large repayments (more than or equal to 1/3; 'coax' condition). The analysis indicates that coaxing elicits larger investments in rounds following coaxing behavior (red) than for rounds following no coaxing (gray). Note that the coaxing-related increase in investment persists across rounds, such that a large repayment in the current round earns larger investment up to 4 rounds into the future.

(C) BPDs coax less. Healthy trustees are twice as likely as BPD trustees to coax when cooperation between players is low. Specifically, healthy trustees are more likely to make a large repayment (≥ investment amount) after having received a small investment (≤ $5). Conversely, BPD trustees are more likely to make a small repayment (< investment amount) after receiving a small investment. The y-axis indicates the proportion of exchanges that trustees repay more than or equal to the investment amount after receiving a small investment (≤ $5).

Fig. 3. Response of 36 healthy trustee brains and 42 BPD trustee brains to level of cooperation.

Top panel: A general linear model (GLM) analysis identified four cortical regions with greater response to small investments relative to large investments (I ≤ $5, n = 109 rounds; I > $10, n = 125 rounds). These regions included bilateral anterior insula (top left), medial frontal gyrus,

bilateral inferior parietal lobule (BA 40), and left inferior frontal gyrus ($p < .001$ uncorrected in yellow; $p < .005$ in orange; see **table S3**). A region-of-interest analysis in the left insula of healthy trustees indicates response in the left insula of trustees scales negatively with the size of investment ($r = -.96$; top right). Percent change in BOLD signal was averaged from the 10 most significant voxels identified in the GLM during the 4-8 second period following the revelation of investment. The mean ± SE of the resulting signal is plotted in $4 bins. In contrast, similar analyses among 42 individuals with BPD showed no increased activation in these regions when investments were small.

Fig. 4. Behavioral and self-report levels of trust indicate perturbed social norms among individuals with BPD.

Left panel: Individuals with BPD report lower trait levels of trust using a self-report measure (Interpersonal Trust Scale; *(61)*). Right panel: Individuals with BPD repay less than the healthy group, indicated lower levels of trust within the exchange game (also see **fig. S2**).