bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The neural circuitry of emotion-induced distortions of trust

Jan B. Engelmann[1,2,3], Friederike Meyer[2], Christian C. Ruff[2,4] & Ernst Fehr[2,4]

[1]Center for Research in Experimental Economics and Political Decision Making (CREED), Amsterdam School of Economics, University of Amsterdam and the Tinbergen Institute
[2]Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, Zurich, Switzerland
[3]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands
[4]Shared senior authorship

Please address correspondence to:

Ernst.fehr@econ.uzh.ch

Jbengelmann@gmail.com

The main article includes:

Figures: 4

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Aversive emotions are likely to be a key source of irrational human decision-making but still little is known about the underlying neural circuitry. Here, we show that aversive emotions distort trust decisions and cause significant changes in the associated neural circuitry. They reduce trust and suppress trust-specific activity in left temporoparietal junction (TPJ). In addition, aversive emotions reduce the functional connectivity between TPJ and emotion-related regions such as the amygdala. We also find that the posterior superior temporal sulcus (pSTS) plays a key role in mediating the impact of aversive emotions on brain-behavior relationships. Functional connectivity of right pSTS with left TPJ not only predicts mean trust taking in the absence of negative emotions, but aversive emotions also largely remove this association between TPJ-pSTS connectivity and behavioral trust. These findings may be useful for a better understanding of the neural circuitry of affective distortions and may thus help identify the neural bases of psychiatric diseases that are associated with emotion-related psychological and behavioral dysfunctions.

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Trust pervades almost every aspect of human social life. It plays a decisive role in families, organizations, markets and in the political sphere. Without trust, families fall apart, organizations are inefficient, market transactions are costly and political leaders lack public support. Recent research in behavioral economics and neuroeconomics has begun to elucidate the determinants and neural correlates of trust [1-3]. However, despite recent progress in understanding the determinants of trust [4] and its distortions in psychiatric disorders [5,6], there are still large gaps in our knowledge about the impact of our emotions on trust taking [7,8], and the underlying neural circuitry. Emotions, in particular those with high intensity, can have deleterious effects on our decision-making faculties, as hinted at by a multitude of public press reports and recent theoretical [9] and experimental accounts [10-13]. It is therefore important to understand the behavioral and neural mechanisms by which emotions distort decisions to trust.

While much progress has been made in outlining the neural underpinnings of emotional processes on the one hand [14,15] and of decision-making on the other [16,17], the effects of emotion on choice have received – with some exceptions [10-12] – less attention to date. Theoretical accounts of the influence of emotion on choice [9] distinguish between two types of emotions: *anticipatory emotions*, such as the anticipated pleasure from the future consumption of a good that reflect how decision-makers expect to feel about the outcomes of their decisions, and *incidental emotions* that occur at the time of the decision, but are unrelated to the choice outcomes. Incidental emotions are of particular interest because of their ubiquity in real life and because they are prime candidates for emotion-induced behavioral distortions. By definition, incidental emotions are unrelated to choice outcomes and, to the extent to which they affect behavior, may cause irrational behavioral biases.

To study the behavioral impact and the underlying neural circuitry of emotion-induced distortions of trust, we adapted the trust game [18] to an imaging context. In the trust game, two anonymous players, which we call investor and trustee, sequentially send money to each other. In the first stage, the investor faces the choice of whether and how much of her endowment to transfer to the trustee. Then the experimenter triples the sent amount, before it is transferred to the trustee. The investor's decision to transfer money thus increases the total amount of money that can be distributed

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

among the two players. In the second stage, the trustee is informed about the total amount that he received after which he can send back part or all of this money. Thus, while the investor's transfer increases the total amount of money available to both parties, the investor also faces the risk of benefitting nothing from the transfer because the trustee is completely free in his back-transfer decision. Therefore, the decision to transfer money constitutes an act of trust, as the investor makes herself vulnerable to the potentially selfish behavior of the trustee [4].

Such trust taking involves both a financial risk due to the possibility of losing the invested money, as well as a social risk of being betrayed by an untrustworthy trustee. Therefore, to enable clear identification of the impact of incidental emotion on the mechanisms involved in the social aspects of trust taking, it is important to include a well-matched non-social control task. For this reason, our subjects also faced a non-social control condition that was identical to the trust condition in every respect except that instead of a trustee, a computer made a "back-transfer" that determined the profitability of the investor's "transfer" [19,20]. The profitability of the investor's transfer in the non-social control and trust condition was exactly the same because the investors had exactly the same choice options in both conditions and the computer sampled the "back-transfer" decisions in the non-social control condition according to the probability distribution of back-transfers that was generated by the trustees' decisions in the trust condition (see online Methods). Therefore, the distinguishing feature between the trust and the non-social control game was the unique possibility of betrayal by the interaction partner in the trust game, which was not present in the non-social control game [2]. This difference between the trust and the non-social control game was saliently indicated at the beginning of each respective trial with either a human-like symbol on the computer screen (in the trust game) or a non-human symbol (in the non-social control game).

Subjects made decisions in either trust or non-social control trials within two different emotional contexts. They were either under the threat of relatively intense tactile stimulation that was somewhat painful ("threat condition"), or they faced the possibility of receiving weak tactile stimulation in the "no threat" condition (Figure 1a). A prolonged period of incidental aversive emotion was established by administering the tactile stimulations at unpredictable time points and frequencies for

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the duration of an entire block. A block consisted of several trust or control trials in the threat condition or several such trials in the no-threat condition. The threat-of-shock paradigm employed in the current study has been shown to reliably induce negative emotion [21-23] and addresses the limitations of standard emotion induction procedures [24,25] as follows: (1) threat of shock provides an immediate stimulus of biological significance that triggers an aversive and automatic emotional reaction, the intensity of which can be measured throughout the experiment using standard psychophysiological techniques; (2) using a blocked threat-of-shock paradigm reinstates the emotional reaction at every presentation, which can thus be maintained for the duration of the entire experiment; (3) threat of shock was administered within-subject, therefore allowing each subject to serve as their own control; (4) tactile stimulation was administered in both the trust and the non-social control condition, thus minimizing demand effects.

In our set-up, decisions to trust entail the unique possibility of being betrayed by the trustee and therefore provide a strong incentive to avoid such betrayal [2]. Therefore, the investor needs to take this aversive outcome into consideration when deciding how much money to entrust the interaction partner, which is accomplished by taking the perspective of the trustee and assess how she will react to given transfers. To identify these trust-specific computations, the non-social control task is identical to the trust game, except that in the latter the investor has to use his social-cognitive abilities to assess how much the trustee – a human being – will send back. This assessment requires mentalizing and perspective taking, in contrast to the non-social control task for which no such processes are necessary to make a decision. Decisions to trust therefore should involve neural processes linked to mentalizing, a hypothesis that is backed up by theoretical accounts [26] and numerous reports that brain areas commonly found to be involved in mentalizing (including DMPFC, STS, TPJ) are activated during the trust game (e.g., [6,27-30]). We conjectured that incidental aversive emotion modulates these trust-specific computations, particularly the simulations of the trustee's reaction to given transfers. This has potentially important implications because if incidental aversive emotion disrupts the recruitment of the social-cognitive processes necessary for mentalizing and perspective taking, we are likely to observe that aversive emotion also has specific effects on the neural mechanisms of trust decisions. In other words, even if aversive emotion would have comparable

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

behavioral effects in the trust game and the non-social control task, we may be able to identify trust-specific effects of aversive emotion on neurocognitive processing because mentalizing and perspective taking are uniquely required in the trust game but not in the non-social control task.

Neurally, we therefore expected that incidental aversive emotion influences trust decisions by specifically modulating neural responses in regions involved in representing other people's mental states, including temporoparietal junction (TPJ) and dorsomedial prefrontal cortex (DMPFC) [31-33]. These regions have been consistently implicated in trust decisions [27,29,30,34,35]. Moreover, they may be affected by incidental aversive emotion because a conjunction of the neurosynth meta-analyses for the terms "emotion" and "theory of mind" identifies an overlap between these networks in TPJ and DMPFC. Together, the above results establish these regions as prime candidates for investigations of the modulatory effects of incidental aversive emotion on mentalizing during trust decisions. In addition, we also expected an effect of the threatening context on areas known to be involved in the processing [14] and regulation of emotions [36], particularly during behavior that requires goal-directed cognition, such as decision-making [37]. A region that meets these criteria is the amygdala, which has consistently [38] and relatively specifically (neurosynth reverse inference analysis for "aversive") been implicated in processing aversive emotion, but also plays a central role in trustworthiness inferences [39,40]. Specifically, we hypothesized that this region is involved in trustworthiness assessments associated with mentalizing during trust decisions in the absence of threat, while in the presence of threat, the amygdala will be preoccupied with evaluating and monitoring the emotionally salient threatening context. We therefore investigated the functional connectivity between regions involved in assessing the trustworthiness of interaction partners during decision-making (TPJ, DMPFC) and regions involved in signaling emotional salience and aversive emotions such as the amygdala.

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Threat of shock induces autonomic arousal and aversive emotions during decision-making

We scanned 41 volunteers while they made trust decisions during the emotionally aversive threat condition and during the emotionally neutral no-threat condition. In a series of manipulation checks, we first assessed whether threat of shock induced emotional arousal by probing galvanic skin conductance responses (SCR), self-reported emotion and brain activations in response to electrical stimulation. As illustrated in Figures 1b and 1c, mean SCR during both trust and non-social control trials were significantly greater during the threat condition compared to the no-threat condition [significant two-way interaction between the factors threat and time: $F(16,624) = 99.28$, $p < 0.001$, $\eta^2 = 0.718$]. Follow-up pairwise comparisons at each time point show significantly enhanced SCR during threat relative to no threat from 2 until 16 seconds after trial onset during trust, and from 3 until 16 seconds after trial onset during non-social control decisions (all two-tailed tests survive Bonferroni correction for multiple comparisons with all $t(39) > 3.313$ and all $p < 0.002$). Taken together, these results indicate significantly greater emotional arousal during the threat condition relative to the no-threat condition in both social and non-social game types.

----------[insert Figure 1]----------

The emotional arousal illustrated in Figures 1b and 1c was clearly experienced as aversive by the subjects. In an open-ended questionnaire administered after scanning, 95.12% of subjects responded that they experienced aversive emotional arousal during threat blocks (Supplementary Figure 1a). The aversive nature of the threat condition was further confirmed by strong activations of central nodes of the brain's pain matrix during the (actual) experience of strong compared to weak tactile shocks (Supplementary Figure 1c) and by the observation of enhanced SCRs following the (actual) experience of strong compared to weak tactile shocks (Supplementary Figure

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1b).

Jointly, the above electrophysiological results, self-reported emotions and activation within the brain's pain matrix during and after the shock, indicate that subjects experienced the threat of a shock as an aversive and arousing emotional state. This state is clearly unrelated to the monetary outcome of trust- and risk-taking, as it does not affect the trustee's or the computer's decisions. The next question we addressed is whether this incidental emotional state distorts subjects' behavior relative to the no-threat control condition.

Aversive emotion reduces investments during trust decisions

To identify whether the aversive emotional state had a significant impact on decision-making, we first investigated mean transfer rates during trust and non-social control decisions for each emotional context and submitted these data to a two-way repeated-measures ANOVA (Mean transfer rates were normally distributed as indicated by the Shapiro-Wilk normality test: $W = 0.976$, $p = 0.510$) with the factors game type (trust, control) and threat (absent, present). Aversive emotional state significantly changed transfers during both trust and non-social control trials (Figure 1d), as indicated by significant main effects of threat [$F(1,40) = 17.483$, $p < 0.001$, $\eta^2 = 0.304$]. Moreover, separate pairwise comparisons (all two-tailed) showed that the threat condition led to a reduction of investments (Figure 1d) in the trust game [$t(40) = -3.4$, $p < 0.005$, mean transfer difference = -1.1 CHF] and in the non-social control game [$t(40) = -3.16$, $p < 0.005$, mean transfer difference = -0.93 CHF]. To exclude the possibility that choices were affected by the *actual experience* of shocks, rather than by the ongoing aversive emotion due to shock expectation, we ran several multiple regression analyses (Supplementary Text 1). The regression results (Supplementary Table 1) show that the behavioral results reported above were indeed due to the aversive emotion ($p < 0.001$) generated by the *threat* of shock, rather than reflecting the effect of actual shock experience immediately before decisions are taken ($p = 0.23$).

Aversive emotion also led to faster reaction times during both trust- and non-social control trials (Figure 1e). Mean reaction times were submitted to a two-way repeated-measures ANOVA (Mean reaction times were normally distributed as indicated by the Shapiro-Wilk normality test: $W = 0.972$, $p = 0.402$) with the factors game type

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(trust, control) and threat (absent, present). We obtained a significant main effect of threat [$F(1,40) = 17.01$, $p < 0.001$, $\eta^2 = 0.298$]. The main effect of threat is characterized by significantly (two-tailed) faster mean reaction times in the threat relative to the no-threat condition (Figure 1e) for both the trust game [$t(40) = -3.3$, $p < 0.005$, mean RT difference = -0.13 s] and the non-social control task [$t(40) = -2.5$, $p < 0.05$, mean RT difference = -0.13 s].

Taken together, these behavioral results indicate that aversive emotion significantly reduced trust taking, as reflected by diminished transfer rates in the trust game. Additionally, aversive emotion reduced transfer rates in the nonsocial control task and reaction times in both the trust and the nonsocial control task. Notably, the absence of a significant interaction between threat and game type for electrophysiological and behavioral measures (transfer rates: $F(1,40) = 0.122$, $p = 0.7$, $\eta^2 = 0.003$, response latencies: $F(1,40) < 0.001$, $p = 0.993$, $\eta^2 < 0.001$, SCR: $F(1,39) = 0.006$, $p = 0.938$, $\eta^2 < 0.001$) indicates that the impact of aversive emotion during trust and nonsocial control trials is similar across these multiple measurement modalities, confirming that our non-social condition constitutes a well-matched control for the trust game.

Aversive emotion suppresses trust-related activity in TPJ

The main goal of our fMRI analyses was to identify the neural circuitry underlying emotion-induced distortions of trust decisions. We therefore first examined brain activation in the ROIs that we conjectured (see our hypotheses in the introductory section) to be preferentially engaged during *trust-specific* computations, such as the assessment of the trustee's trustworthiness and the associated interplay between social cognition and social valuation. Regions involved in representing other people's mental states include temporoparietal junction (TPJ) and dorsomedial prefrontal cortex (DMPFC[41]). We employed small-volume correction at an FWE-corrected threshold of $p < 0.05$ *in truly independent ROIs* defined with reverse inference maps from relevant search terms on neurosynth.org [42] (see online Methods). In the final part of this paper we will also examine the domain general effects (i.e. the effects that are not specific to the trust game) of aversive emotions.

As a first step, our analyses confirmed that several of the conjectured regions were

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

indeed specifically involved in trust (vs. non-social control) decisions when aversive emotion was absent ($\text{Trust}_{\text{no threat}}$ > $\text{NS Control}_{\text{no threat}}$). That is, areas in left TPJ (-57, -60, 27; k = 247; green region in Figure 2a) and DMPFC (-9, 60, 18; k = 39; green region in PFC in Figure 2a) showed significantly greater activation during decision-making in trust relative to control trials in the absence of threat (all activations SV FWE-corrected, see also Supplementary Table 2a). Moreover, we additionally find a large cluster in left TPJ that survives whole-brain correction and extends into Superior Temporal Gyrus (-57, -60, 31, k = 701) and, at a relaxed threshold a cluster in right TPJ ($p < 0.005$, uncorrected; 52, -49, 27, k = 161). We then examined which of the ROIs showed a breakdown of trust-specific activity due to aversive emotion. To identify the threat-induced reduction in brain activation that is specific to the trust game we need to know in which ROI's the contrast ($\text{Trust}_{\text{no threat}}$ > $\text{Trust}_{\text{threat}}$) is significantly *larger* than the contrast ($\text{NS Control}_{\text{no threat}}$ > $\text{NS Control}_{\text{threat}}$). In other words, a positive activation contrast ($\text{Trust}_{\text{no threat}}$ > $\text{Trust}_{\text{threat}}$) is not sufficient for a trust-specific effect because it could also be the case that in the non-social control condition the threat of shock induces a reduction in brain activation. Therefore, we examined the breakdown of trust-specific activity due to aversive emotion by computing the following interaction contrast: ($\text{Trust}_{\text{no threat}}$ > $\text{Trust}_{\text{threat}}$) > ($\text{NS Control}_{\text{no threat}}$ > $\text{NS Control}_{\text{threat}}$). Note that this interaction contrast is orthogonal to the above mentioned simple contrast ($\text{Trust}_{\text{no threat}}$ > $\text{NS Control}_{\text{no threat}}$) because the sum of the products of the corresponding coefficients for the simple effects and interaction contrasts adds to zero.

A region in left TPJ indeed showed a significant interaction effect (-60, -54, 19; k = 95, SV FWE-corrected, yellow region in Figure 2a, see also Supplementary Table 2b). Note, however, that our findings of trust-specific suppression in left TPJ also extend to right TPJ at a relaxed threshold ($p < 0.005$, uncorrected; 51, -48, 31, k = 88). To further characterize this interaction effect we examined post hoc the impact of aversive affect in the trust game separately from its impact in the non-social control game: the results of the contrast ($\text{Trust}_{\text{no threat}}$ > $\text{Trust}_{\text{threat}}$) during trust trials shows a suppression of activation for trust decisions within left TPJ (-58, -55, 19; k = 103; SV FWE-corrected, Supplementary Table 2c). During non-social control decisions on the other hand, no voxels in our left TPJ ROI showed greater activity during no threat relative to threat ($\text{NSC}_{\text{No Threat}}$ > $\text{NSC}_{\text{Threat}}$), even at a very liberal threshold of $p < 0.05$,

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

uncorrected (see also Supplementary Figure 2a for additional univariate analyses that underline the strength of the interaction effect in left TPJ). These results indicate that the interaction effect is based on a selective interference of aversive emotion with trust-related activity, but not with activity in the nonsocial control condition.

----------[insert Figure 2]----------

Given that decisions involving trust rely on neural circuitry that mediates the interplay between social cognition and valuation [1], we also performed an exploratory analysis of the impact of aversive emotion on trust-related activity within regions commonly implicated in valuation (vmPFC and ventral striatum [43]). These results show reductions in trust-related activity due to aversive emotion in vmPFC and ventral striatum (Supplementary Table 2c). For completeness, we also conducted a whole brain analysis to identify potentially important activations outside our a-priori regions of interest. The whole brain analysis shows that in the left superior temporal sulcus, posterior cingulate cortex and left inferior parietal lobe (Supplementary Figure 3 and Supplementary Table 3) aversive emotion suppresses brain activity only in the trust game but not in the nonsocial control game. However, in these regions we do not find a significant interaction effect in the sense that suppression of activity due to aversive emotion is significantly stronger in the trust relative to the nonsocial control task.

Aversive emotion suppresses trust-specific connectivity between TPJ and amygdala

Recent studies stress the importance of the interplay between cognitive and emotional networks [14,37]. Therefore, we investigated the effects of aversive emotion on the connectivity between trust-relevant brain regions with Psychophysiological Interaction analyses (PPI) [44]. In view of the key role of the temporoparietal junction (TPJ) in perspective taking and mentalizing [32,45] and the conjecture that these mental operations are important for trust taking, and our finding of enhanced activity in TPJ in the trust compared to the control task (see above), we were particularly interested in how aversive emotion changes the functional connectivity between the TPJ and

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

emotion processing regions, such as the amygdala. ROI analysis of the data during the *threat-absent* condition ($\text{Trust}_{\text{no threat}}$ > NS $\text{Control}_{\text{no threat}}$) confirmed that the connectivity between the TPJ and a region in left amygdala (-22, -9, -15; $p < 0.05$, SV FWE-corrected, k = 14) is indeed stronger during decisions in the trust relative to the control task (Supplementary Table 4a). Moreover, at a relaxed threshold a cluster in right amygdala also shows stronger connectivity with TPJ during trust compared to non-social decision-making ($p < 0.005$, uncorrected; 22, -6, -11, k = 25). Therefore, we were interested whether there is a *trust-specific* threat-induced connectivity change. To answer this question we performed an interaction analysis that examined whether the threat-induced connectivity change in the trust condition, ($\text{Trust}_{\text{no threat}}$ > $\text{Trust}_{\text{threat}}$), is larger than the threat-induced connectivity change in the control condition (NS $\text{Control}_{\text{no threat}}$ > NS $\text{Control}_{\text{threat}}$). Note that this interaction contrast is orthogonal to the above mentioned simple contrast ($\text{Trust}_{\text{no threat}}$ > NS $\text{Control}_{\text{no threat}}$) because the sum of the products of the corresponding contrast coefficients adds to zero.

This analysis revealed threat-induced aversive emotion causes a stronger connectivity change between TPJ and a region in the amygdala [left: -26, 0, -23; $p < 0.05$, SV FWE-corrected, k = 29, Figure 2c shown in yellow; right at relaxed threshold of $p < 0.005$, uncorrected: 28, -7, -21, k = 18] in the trust game compared to the control task (see Supplementary Table 4b). We performed a post-hoc inspection of the significant interaction in left amygdala by investigating the effect of threat on connectivity changes for the trust and the non-social control condition separately. We find that aversive emotion disrupted functional connectivity specifically during trust (compare red vs green bars in Figure 2d) but not during non-social control decisions. A follow-up contrast investigating threat effects on trust-related connectivity patterns ($\text{Trust}_{\text{No threat}}$ > $\text{Trust}_{\text{Threat}}$) confirmed the suppression of TPJ-amygdala connectivity during trust decisions (-28, -6, -14; $p < 0.05$, SV FWE-corrected, k = 110, Supplementary Table 4c). In contrast, during decisions in the non-social control task no voxels in our left amygdala ROI showed greater connectivity during no threat relative to threat, or the reverse contrast of threat vs. no threat, even at a very liberal threshold of $p < 0.05$, uncorrected (see also Supplementary Figure 2b for additional analyses that underline the strength of the interaction effect in left amygdala). Moreover, comparison of connectivity during decisions in the control compared to the trust task in the threat

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

condition shows significant suppression of connectivity during trust relative to non-social decisions in left amygdala (-26, 0, -23, $p < 0.05$, SV FWE-corrected, k = 14), indicating that the threat-related suppression of TPJ-amygdala connectivity is also evident when comparing NS control to trust. Together, these results indicate that threat causes specific suppression of connectivity during trust taking that can be observed when comparing the effect of threat within the trust task (Trust: no threat > threat), as well as the effect of threat on connectivity during trust relative to NS control decisions (Threat: ns control > trust). This suppression of TPJ-amygdala connectivity during trust decisions occurs in the absence of suppression during non-social control decisions (ns control: threat = no threat). Thus, aversive emotion not only affected trust-specific overall activation in the TPJ, but also led to trust-specific connectivity changes of this area with the amygdala.

A trust network: TPJ connectivity strength with pSTS, DMPFC and VLPFC specifically predicts behavioral trust

The above PPI analyses show the average impact of aversive emotion on the functional connectivity between TPJ and amygdala. However, as we observed strong individual differences in the functional connectivity between TPJ and amygdala on the one hand, and in mean transfer levels on the other hand, we next asked the question how individual differences in functional TPJ connectivity are related to individuals' mean transfer levels in the absence and the presence of aversive emotion. Following our analysis approach for TPJ activity and connectivity above, we first identify trust-specific brain-behavior correlations by comparing the relationship between transfer rates and TPJ functional connectivity as a function of game type in the absence of threat ($\text{Trust}_{\text{no threat}}$ > NS $\text{Control}_{\text{no threat}}$). Specifically, we examined whether in our a priori ROIs the relationship between mean transfers and functional TPJ connectivity is different in the trust game compared to the non-social control game via a flexible factorial model that, in addition to the factors Subject, Task and Threat, also includes mean transfer levels in each condition as covariates. We find that in the conjectured ROIs the connectivity between the seed region in the left TPJ and the right amygdala [27, 2, -20, k = 90], the DMPFC [-8, 40, 20, k = 692] and the right STS [64, -43, 4, k = 157; note (see online methods) that the STS is the most

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ventral part of our TPJ mask] exhibits a significantly stronger positive correlation with mean transfer rates in the trust game compared to the non-social control game (all $p < 0.05$, SV FWE-corrected, Supplementary Table 5a). In the next step, we conducted an exploratory whole-brain analysis (FWE-corrected at the cluster level) that identified an extended network of regions (Figure 3a-b, d) and Supplementary Table 6a) that showed a difference in the relationship between individuals' mean transfers and their functional TPJ connectivity across trust and control tasks (in the absence of threat). The regions in this network comprised DMPFC [-2, 17, 54, k = 6216], the right superior temporal sulcus (STS) extending into angular gyrus [right: 64, -43, 4, k = 1188, left: at a relaxed threshold of $p < 0.005$, uncorrected, -62, -45, 4, k = 22] and other regions such as bilateral ventrolateral PFC (see Supplementary Table 6a]. In all these regions we observe a positive and significantly stronger correlation between mean transfer levels and functional TPJ connectivity in the trust compared to the control task (see Figure 3a-b, d and Supplementary Table 6a).

Finally, we also tested whether the slightly negative slopes observed in the regression lines connecting TPJ connectivity and mean transfer rates in the non-social control conditions of Figure 3a-b and 3d are statistically significant. To this end, we ran simple effects contrasts probing for a correlation between TPJ connectivity strength and mean transfer during NS control decisions in the absence of threat. We found no evidence that TPJ connectivity with its target regions negatively predicts transfer rates in the NS control condition, even at a relaxed threshold of $p < 0.05$.

Taken together, the above results therefore confirm the conjecture that there is a trust-specific functional connectivity between the TPJ and amygdala that is suppressed by aversive emotions. In addition, the larger the TPJ connectivity with key regions implicated in mentalizing (the DMPFC and right STS) and emotion (the amygdala), the more subjects are willing to trust their partners on average. Moreover, this predictive relationship between transfer rates and TPJ connectivity is absent in the non-social control game. These results thus suggest a trust-specific network based on TPJ activity and the connectivity of TPJ with a network consisting of the amygdala, right STS, DMPFC and bilateral VLPFC.

----------[insert Figure 3]----------

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Aversive emotion removes the relationship between TPJ connectivity strength and behavioral trust

How does the relationship between functional connectivity patterns in the trust network and behavioral trust-taking change if subjects are exposed to aversive emotion? To answer this question, we investigated whether there is a specific breakdown of the association between mean transfer rate and TPJ connectivity during trust relative to control decisions in the presence of threat via the interaction contrast ($\text{Trust}_{\text{no threat}} > \text{Trust}_{\text{threat}}$) > ($\text{NS Control}_{\text{no threat}} > \text{NS Control}_{\text{threat}}$). We first test for the presence of this effect in our a priori ROIs in TPJ and DMPFC. We find a significant effect in the right STS [64, -43, 6, $p < 0.05$, SV FWE-corrected, $k = 19$; Supplementary Table 5b, recall (see online methods) that the STS is the most ventral part of our TPJ mask]. We then explored whether other regions also show this interaction via a whole brain analysis and find that only the right STS shows an interaction effect (64, -43, 6, $k = 482$; Supplementary Table 6b). To characterize the interaction effect, we ran post hoc simple effects analyses that compare the relationship between transfer rates and TPJ functional connectivity as a function of threat in the trust game. Specifically, we examined whether aversive emotion caused significant changes in the relationship between TPJ connectivity (with its target regions) and mean trust levels in the trust game. This analysis showed that aversive emotion caused a general breakdown of the association between left TPJ connectivity and mean trust taking in the right posterior superior temporal sulcus [64, -43, 6, $k = 698$, Supplementary Table 6c], as well as two other regions mentioned in supplementary table 6. ROI analyses confirm this effect in STS [64, -43, 6, $p < 0.05$ SV FWE-corrected, $k = 55$; Supplementary Table 5c]. In these regions, therefore, there is a significantly positive relationship between left TPJ connectivity and mean trust levels during the no-threat condition (Figure 3c, green regression line) that vanishes in the presence of threat (Figure 3c, red regression line). In contrast, during decisions in the non-social control task no voxels in any of our ROIs, as well as the superior temporal sulcus showed greater connectivity during no threat relative to threat, or the reverse contrast of threat vs. no threat, even at a very liberal threshold of $p < 0.05$, uncorrected.

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Please note that the region in pSTS that is affected in this way by aversive emotions significantly overlaps with the region that shows stronger TPJ connectivity in the trust relative to the control task (overlapping region is shown in yellow in Figure 3). These results suggest that connectivity between left TPJ and its target region in contralateral pSTS supports general trust taking when distortionary aversive emotion is absent. However, in the presence of aversive emotion the relationship between the connectivity pattern in the trust network and behavioral trust taking reverses. Thus, aversive emotion not only reduces average trust taking, but it also diminishes specific relationships between the connectivity patterns in the trust network and behavioral trust taking. Our results therefore suggest that the pSTS is a crucial neural node that mediates the breakdown of trust in the presence of threat.

Aversive emotion alters activation patterns within choice-relevant domain-general neural circuitry

The previous analyses indicate that aversive emotion had distinct effects on neural processes devoted to trust decisions and that functional connectivity strength between TPJ and its targets was specifically related to behavioral trust. However, the pronounced emotional reactions to the threatening context also had an impact on non-social control decisions. We therefore addressed the question to what extent aversive emotion impacts general choice-related neural circuitry in both the trust and the non-social control game by investigating the main effect of aversive emotion: ($\text{Trust}_{\text{threat}}$ + NS $\text{Control}_{\text{threat}}$) > ($\text{Trust}_{\text{no threat}}$ + NS $\text{Control}_{\text{no threat}}$). We identified a domain-general network of regions that show suppression and enhancement in choice-related neural activity during both the trust and the non-social control task (Figure 4 and Supplementary Table 7). The suppression of neural activity in the threat condition (red time course, Figure 4) relative to the no threat condition (green time course, Figure 4) is observed in both the trust and non-social control task (Supplementary Figure 4) in bilateral posterior dlPFC [left: -62, -4, 18, k = 1901 and right: 62, -6, 28, k = 1010], left amygdala [-24, -15, -23, k = 552], posterior paracentral lobule [4 -36, 69, k = 887], and a large cluster (k = 4082) that includes left vlPFC [-48, 41, -8] and vmPFC [-10, 44, -8; Supplementary Table 7a]. Significant enhancement of activity during decision-making under aversive emotion (Supplementary Figure 5) was obtained in the thalamus [18, -6, 1, k = 559] and cerebellum [-4, -46, -24, k = 849;

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 7b]. Together, these results identify a network of domain-general regions, whose decision-related activity is significantly impacted by incidental aversive emotion. Notably, the regions identified by the main effect do not overlap with regions showing trust-specific effects (tested via conjunction analysis), underlining that the trust-specific effects of aversive emotion occur above and beyond domain-general effects on decision-making.

----------[insert Figure 4]----------

Discussion

Incidental aversive emotion is a ubiquitous phenomenon that pervades many aspects of human behavior and human social interaction. In this paper, we investigated the behavioral and neural impact of incidental emotion on trust decisions. We employed a novel experimental technique to establish aversive emotion by inducing a prolonged expectation of unpredictable and aversive tactile stimulation embedded within a hybrid fMRI design. The threat of painful tactile stimulation significantly increased autonomic arousal during both social and non-social decision-making and was associated with consistent self-reports of the experience of aversive emotion. We observed that aversive emotion significantly reduced subjects' trust in their partners. To the extent to which aversive emotions are associated with stress, this result is consistent with a recent behavioral study that showed that acute stress reduces trust [7]. Importantly, despite the fact that aversive emotion was incidental to the decisions made by subjects, we observed significant behavioral and neural effects of the aversive emotional state. The behavioral impact of incidental aversive emotion contradicts consequentialist economic models that assume that emotions can at most affect choices by changing subjects' preferences over outcomes[9]; our results thus also underline the importance of emotions in decision-making, even when they are unrelated to choice outcomes.

Our neuroimaging results provide information about the neural mechanisms behind the reduction of participants' trust. We show a disruption of trust-specific neural activity and connectivity due to aversive emotion. While the dorsomedial PFC

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(DMPFC) and, most notably, the temporoparietal junction (TPJ), were preferentially engaged during trust decisions, aversive emotion led to a trust-specific breakdown of this activation pattern (Figures 2a-b) in the left TPJ and in right TPJ (at a reduced threshold). Moreover, in the absence of aversive emotion, we observed significant connectivity between TPJ and amygdala during trust taking, but this connectivity was disrupted when we induced aversive emotion (Figures 2c-d). Aversive emotion also disrupted the relation between the connectivity patterns in the neural network underlying trust and the extent of behavioral trust-taking. In particular, functional connectivity strength with the TPJ predicted mean trust levels for a network of regions consisting of amygdala, DMPFC, superior temporal sulcus (STS), and VLPFC (Supplementary Tables 5 and 6, Figures 3a-b, d). Aversive emotion caused a specific breakdown of this association for the posterior STS, such that the connectivity between left TPJ and right pSTS no longer predicted overall trust-taking (Figure 3c). Our results therefore identify a network of interconnected regions consisting of left TPJ, amygdala and right pSTS, for which connectivity patterns during trust-taking are significantly impacted by incidental aversive emotion.

The previous literature [3,46,47] has identified betrayal aversion as one of the key determinants of trust-taking in the trust game. Betrayal aversion means that subjects find it extremely aversive to be cheated in the trust game by an untrustworthy partner. Betrayal aversion therefore constitutes a powerful motivation to form expectations about the partner's responses to the various trust levels and to assess the emotional significance of these responses. These processes critically involve subjects' mentalizing faculties and the assessment of the (negative) emotional value of the possibility of being cheated by an untrustworthy partner [20,27,29,48-50].

The temporoparietal junction (TPJ) and the DMPFC have repeatedly been implicated in representing and interpreting others' mental states and behavior [31,32,41,51-53] and the amygdala has been shown to respond strongly to variations in the trustworthiness of faces [39,40] and other emotionally salient stimuli [54,55]. We therefore conjectured that these regions also play a key role in the computations involved in assessing and evaluation the partner's anticipated responses in the trust game. This hypothesis is consistent with prior reports of TPJ, DMPFC and amydgala activation during trust decisions [20,27,29,34,35,48]. Our present results significantly extend these prior findings, by

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

underlining the behavioral relevance of interacting neural networks (rather than just isolated areas) for trust decisions. This is most consistently shown by our findings that enhanced connectivity between TPJ and regions important for social cognition (DMPFC, STS) and emotion processing (amygdala, VLPFC) relate to each individuals' levels of trust taking.

Most importantly, however, we show that key components of these trust-supportive and trust-specific neural networks are suppressed by aversive emotion. In particular, we find that threat of shock leads to (1) specific reductions of TPJ activation during trust decisions; (2) specific reductions in the connectivity between TPJ and amygdala during trust decisions; (3) general reductions in choice-related activity in the amygdala; and (4) trust-specific disruptions of the association between TPJ connectivity and mean trust taking. These results thus suggest that aversive emotions undermine decisive components of trust-specific neural networks involved in the computations relevant for assessing a partner's responses to various trust levels and the associated emotional valuations. This effect is expressed particularly clearly in the change in TPJ-amygdala connectivity due to threat-of-shock: In the absence of threat, TPJ and amygdala show trust-specific communication likely reflecting the integration of social cognitive (mentalizing, TPJ) and social emotional (trustworthiness assessments and evaluations, amygdala) information important for trust decisions; in the presence of threat, by contrast, the amygdala shows a general suppression due to the emotional context (Supplementary Figure 4) and this preoccupation with the immediate and emotionally highly salient threatening context seems to prevent it from communicating with TPJ. In other words, aversive emotions seem to reduce a subject's capacity to mentalize about and evaluate the emotional consequences of various trust levels and, as a consequence, subjects reduce their behavioral trust towards their partner.

In conclusion, we report results that show a significant behavioral impact of incidental emotions on trust-taking and we identify the trust-specific neural mechanisms associated with the impact of aversive emotion on trust taking. These effects were observed even though induced emotions were unrelated to the choice outcomes in our task, confirming that incidental emotions can have a powerful impact on behavior and its underlying mental operations. Our findings inform the development of economic

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and social theory and call for the integration of incidental emotion in behavioral models of social and non-social decision-making [9,56,57]. In addition, by identifying the specific distortions of the neural network activity supporting trust taking, we provide a first step towards neural models that help us better understand such distortions. In particular, our results support the notion that an important mechanism through which aversive incidental emotion impacts social decision-making is the suppression of activity and connectivity between regions known to be crucial for mentalizing about other people's responses (such as the temporoparietal junction, dorsomedial PFC and the superior temporal sulcus) and the evaluation of socially threatening stimuli (such as the amygdala). Given that psychiatric diseases, such as pathological anxiety, social phobia or depression, are characterized by a particularly pronounced susceptibility to negative emotion, our results may also be useful in understanding the neural circuitry associated with emotion-related distortions of social behavior in psychiatric diseases.

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Participants

41 human volunteers (mean age (std.) = 22 (2.145), 17 females) from various Universities in Zurich participated in the current experiment. Only right-handed subjects between the ages of 18 – 45 with no prior psychiatric illness, no regular illicit drug use and no traumatic head injury were included in the experiment. The sample size was based on recommendations for investigations of individual difference in fMRI studies based on power simulation [58]. All participants gave written informed consent to procedures approved by the local ethics committee (Kantonale Ethikkommission, Zurich, Switzerland) before participating in the study. Subjects were right-handed as assessed by the Edinburgh handedness questionnaire and did not report any history of psychological illness or neurological disorders, as assessed by a standard screening form.

Prescanning procedure

Particular care was taken to ensure that subjects understood all aspects of the experiment. To this end, subjects were instructed to carefully read detailed instructions and were required to fill out an extensive questionnaire probing their understanding of the experimental procedures. The accuracy of each subject's answers was confirmed by the experimenters and discussed in a brief interview that lasted for ca. 10-minutes. Subjects were then placed inside the scanner for a brief practice session consisting of 12 trials to ensure that they could view all stimuli, perform the task, make decisions in the allotted 5.5 seconds per trial, understood the experimental setup and to give subjects the opportunity to ask further questions.

After completion of practice, subjects were taken out of the scanner and washed their hands before placement of SCR and stimulation electrodes. Subjects were then placed inside the scanner and two ring electrodes were attached to the dorsum of the left hand: (1) the electrode providing relatively higher intensity stimulation was placed between one to two cm below the second carpometacarpal joint, and (2) the electrode providing relatively lower intensity stimulation was placed one to two cm below the

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fifth carpometacarpal joint. To determine individual thresholds for high-intensity and low-intensity stimulation, we followed a standard procedure [59,60] and employed a visual analog rating scale (VAS) with endpoints defined as 0 = 'cannot feel anything' and 10 = 'maximum tolerable pain'. Tactile stimulation was delivered via two Digitimer DS5 isolated bipolar constant current stimulators (bipolar constant current, 5V, 50mA, Digitimer Ltd, Welwyn Garden City, UK) and a custom-made fMRI compatible 5-mm ring electrode, which delivered a maximally focused and centered tactile stimulus. By varying current amplitude between 1 and 99 % of maximum amperage, stimuli with varying intensity levels were repeatedly delivered to each participant until stable ratings were achieved at least three times according to the following criteria: between 1 and 2 for the low intensity stimulus, and between 8 and 9 for the high intensity stimulus. Visual and tactile stimulus presentation, as well as recording of responses, were controlled by Cogent2000 (http://www.vislab.ucl.ac.uk/cogent.php).

Task

To investigate the effect of incidental emotion on trust-taking, we employed a hybrid fMRI design, in which aversive emotion was manipulated in a blocked fashion while social (trust) and non-social (control) tasks were presented in an event-related fashion. Specifically, we varied aversive emotion by creating an expectancy of weak or strong unpredictable electrical stimulation that could occur at any time for the duration of an entire block. This expectancy was created by means of a block cue presented at the beginning of each block that informed participants about the game type (trust or control game) and the intensity of stimulation (weak or strong) for the current block (Figure 1a). Stimulation intensity was communicated to subjects in three ways: (1) via a verbal cue embedded in the 750-ms block cue ["strong" for treatment ("threat" condition), "weak" for control ("no-threat" condition)]; (2) via a tactile reminder cue presented 700 ms after visual cue onset that reflected the exact stimulation intensity of the current block; (3) via a specific background color that was consistently associated with either threat or no-threat blocks for each subject (color was counterbalanced across subjects) and remained constant for the duration of a block. The number and time points of electrical stimulation events throughout the blocks were determined to be completely unpredictable to subjects, in order to augment the efficacy of the threat-

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of-shock treatment. For this purpose, the number of stimulation events was determined for each block by random draw from a gamma distribution (shape parameter = 1; scale parameter = 1). The exact timing of these stimulation events was then determined at random time points between the offset of the cue display and onset of the resting screen drawing from a uniform distribution, with the constraint that at least 0.2 s separated successive electrical shocks. Timing and order of stimuli were randomized for each subject to maximize identification of the effects of aversive emotion on the neural correlates of trust decisions using in-house software programmed in Matlab.

Each block commenced with the set of cues described above that indicated the type of decision to be made (non-social control or trust) and the level of stimulation (weak, strong) to be expected by subjects for the rest of the block. After a brief and jittered interstimulus interval of 3-9 seconds, the first of three trials within a given block was displayed. In both the trust and the control game, subjects were presented with a multiple-choice scenario, in which one of five amounts between 0 and 24 Swiss Francs (CHF) could be transferred to Player B or invested in a lottery. While subjects always had the options to either invest all (24 CHF) or none (0 CHF) of their endowment, each trial presented a novel choice scenario by (1) varying the intermediate options between 4, 6, or 8 CHF in the low category, 10, 12, or 14 CHF in the medium category, and 16, 18, or 20 in the high category of intermediate transfer amounts; (2) varying the location of each choice option and (3) varying the location of the originally highlighted choice option. This variability was introduced in order to ensure that subjects paid attention to all choice options on every trial and to avoid excessive use of heuristics. Intermediate amounts, location of choice options and location of the initially highlighted choice option were fully counterbalanced across conditions. Subjects selected their preferred option by moving a yellow dot that highlighted the currently selected choice option up and down by pressing two dedicated buttons on a standard MR-compatible 4-button response box and confirming their choice by pressing a third button. At this point the selected choice option was highlighted in red for the remaining duration of a trial. After a jittered intertrial interval (3-9 seconds) a new trial began. Please note that in order to control for wealth effects, subjects in our experiment did not receive trial-by-trial feedback about the

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

financial outcome of their choices in both the trust and the non-social control game. By using one-shot games with no feedback we preclude learning- and outcome- related signals commonly observed in valuation regions (e.g., [51,61]). Subjects completed 28 blocks (7 blocks per condition with an average length of 38.75 seconds) with three trials each in two runs.

Payment determination

We collected trustee responses in separate behavioral sessions that were conducted prior to the fMRI experiment using the same trust game. We elicited the trustees' choices with the strategy method, i.e. the trustees indicated their responses to each feasible transfer level. The trustees gave written and informed consent that we could use their strategies in follow-up experiments. In the fMRI part of our experiment the subjects (investors) thus played against the pre-recorded strategies of the trustees, i.e. a subject's transfer level together with the strategy of the (randomly) matched trustee determined the final monetary outcome in a trust game trial. Given the absence of the trustee on the scanning day, we informed participants that they were interacting with trustees in a temporally delayed fashion. Specifically, we emphasized to subjects that their payoffs were determined by decisions of real persons in the trust game, and by a computer algorithm in the control game, and that they were assigned different real persons across trust game trials. Finally, to maintain the interactive nature of the trust game, we informed our subjects that their choices had real, but delayed, consequences for trustees, who were sent additional payments according to the decisions made by the investor in the scanner after completion of the experiment. During the experiment, the subjects did not receive any feedback about the behavior of their matched trustees, or the payoff amounts from lottery investments.

After completion of the fMRI part of the experiment, the subjects selected two trials at random by dice throw and payment was determined according to the decisions made by the subject and the trustee on the selected trials. In order to avoid hedging, both payout trials were drawn from the entire experiment, i.e. the payout trials were not specific to a condition, such as the trust or control game. If a trust game was randomly chosen for payout determination, the investor's payout was determined based on the amount transferred to the trustee and the backtransfer amount of the

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

specific trustee the investor was paired with on that trial (payout investor = 24 – transfer to trustee + backtransfer from trustee; payout trustee = 24 + transfer from investor * 3 – backtransfer to investor). If a control game was randomly chosen, the computer algorithm randomly drew a payout amount from the distribution of trustees' backtransfer amounts. Our procedure therefore created equivalent payout amounts and likelihoods for the trust and control game.

Exit questionnaire

After completion of the experiment, subjects filled out an exit questionnaire that probed their beliefs about the accuracy of our instructions, as well as emotional reactions to our experimental manipulations. The main goal of the exit questionnaire was to measure whether subjects believed our instructions. Note that we implement such measurements routinely although we have little reason to believe that subjects doubt our instructions. Our laboratory uses deception only as a very rare exception, and we also did not use any deception in this experiment and fully disclosed all information truthfully to the subjects. Subjects were asked to rate 7 statements on a scale from 0, indicating very unbelievable, to 4, indicating very believable. The statements declared that the trust games were played with real persons, that each trust game was played with an anonymous trustee, that decisions of trustees were made by actual persons, and that trustees will receive additional payments based on the decisions of subjects on the relevant trial. Subjects' responses were entered into one-sample t-tests testing whether responses were significantly greater than the mid-point of the scale (2, indicating neither believable nor unbelievable). Mean ratings for all statements were significantly greater than two, indicating that subjects believed the statements (all t-tests survive the Bonferroni-corrected threshold of 0.007; average rating (SD) over all statements is 3.37 (0.86)).

Skin conductance responses (SCRs)

Skin conductance responses were collected using a PowerLab 4/25T amplifier with a GSR Amp (ML116) unit and a pair of MR-compatible finger electrodes (MLT117F), which were attached to the participants' left middle and ring finger via dedicated Velcro straps after application of conductance gel. Subjects' hands had been washed

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

using soap without detergents before the experiment. Stable recordings were ensured before starting the main experiment by waiting for signal stabilization during training and stimulation intensity calibration. LabChart (v. 5.5) software was used for recordings, with the recording range set to 40 µS and using initial baseline correction ("subject zeroing") to subtract the participant's absolute level of electrodermal activity from all recordings (all specs for devices and electrodes from ADInstruments Inc., Sydney, Australia).

Due to technical problems, data from 4 subjects included 1 run (out of 2) and data from 1 subject was lost. Each participant's SCR data were initially smoothed with a running average over 500 samples (equivalent to 500 ms at a sampling rate of 1KHz) to reduce scanner-induced noise. Data were then resampled from 1KHz to 1Hz and subsequently z-transformed. Statistical analysis of the pre-processed skin conductance data followed the approach commonly employed in analyses of fMRI data. Specifically, multiple linear regression implemented in AFNI was used to estimate SCR during decisions made in each of the task conditions, that is, during trust and non-social control tasks and in the context of threat and no-threat treatment blocks. The statistical model included a total of 7 regressors that reflected the onset times of decision screens in trust and non-social control trials under expectancy of strong and weak electrical shocks, cue times indicating the onset of a block, as well as delivery times of strong and weak tactile stimulation. To avoid making assumptions about the shape of the SCR response, a finite impulse response (FIR) model was used to estimate average responses (beta weights) during each trial type via deconvolution from event onset to 16s post onset using 17 cubic spline basis functions. Constant, linear and quadratic terms were included as regressors of no interest for each run separately to model baseline drifts of the SCR. Regressor estimates (beta weights) at each time point and for each condition were then used in follow-up analyses reported in the Results section.

fMRI data acquisition

Magnetic resonance images were collected using a 3T Philips Intera whole-body magnetic resonance scanner (Philips Medical Systems, Best, The Netherlands) equipped with an 8-channel Philips sensitivity-encoded (SENSE) head coil. Structural

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

image acquisition consisted of 180 T1-weighted transversal images (0.75-mm slice thickness). For functional imaging, a total of 1095 volumes were obtained using a SENSE T2*-weighted echo-planar imaging sequence [62] with an acceleration factor of 2.0. We acquired 45 axial slices covering the whole brain with a slice thickness of 2.8 mm (inter-slice gap of 0.8 mm, sequential acquisition, repetition time = 2470 ms, echo time = 30 ms, flip angle = 82°, field of view = 192 mm, matrix size = 68 × 68). To optimize functional sensitivity in orbitofrontal cortex and medial temporal lobes, we used a tilted acquisition in an oblique orientation at 15° relative to the AC-PC line.

fMRI data analysis

Preprocessing and statistical analyses were performed using SPM8 (Wellcome Department of Imaging Neuroscience, London, UK). To correct for head motion, all functional volumes were realigned to the first volume using septic b-spline interpolation and subsequently unwarped to remove residual movement-related variance due to susceptibility-by-movement interactions. Slice timing correction was performed after realignment/unwarping. To improve coregistration, bias-corrected anatomical and mean EPI images were created and subsequently coregistered using the new segment toolbox in SPM. Images were normalized to the Montreal Neurological Institute T1 template using the parameters (forward deformation fields) derived from the nonlinear normalization of individual gray matter tissue probability maps. Finally, functional data underwent spatial smoothing using an isotropic 6-mm FWHM Gaussian kernel.

Statistical analyses were carried out using the general linear model. Regressors of interest were modeled using a canonical hemodynamic response function (HRF) with time and dispersion derivatives in order to account for subject-to-subject and voxel-to-voxel variation in response peak and dispersion [63]. Since our main interest was the impact of aversive emotion on trust-taking, we modeled the decision period for the full response time on each trial, that is from the onset of the decision screen until subjects pressed the confirm button. This was done in the following four conditions: (1) trust game during relatively high-intensity stimulation expectancy (threat condition), (2) trust game during relatively low-intensity stimulation expectancy (no-threat condition), (3) control game during relatively high-intensity stimulation

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expectancy (threat condition) and (4) control game during relatively low-intensity stimulation expectancy (no-threat condition). Finally, the following regressors of no interest were included in our model: the actually realized weak and strong tactile stimulations during a block (one reminder shock and, on average, one additional shock randomly drawn from a gamma distribution were administered per block), block cues indicating game type (trust, control) and stimulation intensity of the reminder shock (weak, strong) at the beginning of each block, as well as omissions of behavioral responses during a trial.

The main goal of the current investigation was to identify the impact of aversive emotion on the neural correlates of trust decisions. Trust-specific neural effects of aversive emotion can be identified via an interaction between threat and game type, in which threat significantly alters the neural correlates of decision-making in trust relative to non-social control trials. To investigate the interaction between threat and game type, an ANOVA was computed by entering contrast estimates obtained from first level models into a flexible factorial model with the factors game type (trust, control), threat (absent, present), as well as subject. We were particularly interested in trust-specific emotion-induced suppression of activity and connectivity, which we tested via the interaction contrast ($\text{Trust}_{\text{no threat}} > \text{Trust}_{\text{threat}}$) > ($\text{NS Control}_{\text{no threat}} > \text{NS Control}_{\text{threat}}$) in the context of the flexible factorial design. A covariate reflective of each subject's mean transfer in each condition was also included to probe for brain-behavior correlations. All analyses were also conducted without the behavioral covariate and results did not change.

We expected regions commonly implicated in the major cognitive and affective component processes of trust taking to be affected by aversive emotion. Specifically, we hypothesized that subjects needed to assess the trustworthiness of the trustee to make predictions about payout probability in the trust game, which involves regions commonly implicated in theory of mind and social cognition [64,65]. To identify regions implicated in theory of mind, we consulted neurosynth.org [42], which offers a means to obtain automated meta-analyses over a large number of prior fMRI investigations and thereby provides an independent method to obtain masks for ROI analyses. To guide and constrain our ROI selection, we computed the conjunction of the neurosynth meta-analyses for the terms "emotion" (forward inference to identify regions that

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consistently show modified activity) and "theory of mind" (reverse inference to identify regions that are *specifically* involved in theory of mind). This approach identified overlap between these networks in left TPJ and DMPFC, which agrees particularly well with results from recent a meta-analysis identifying the TPJ and DMPFC as core social cognition regions[45]. Furthermore, because of its prominent role in signaling emotional salience [14] , we included the amygdala as an additional ROI (see also neurosynth search: "emotion").

ROI analyses in relevant cortical regions were conducted using small volume correction with masks created via relevant search terms on neurosynth.org, while anatomically well-defined subcortical ROI masks were created using the AAL atlas implemented in WFU Pickatlas. The following independent ROI masks were created via automated meta-analyses from neurosynth.org: (1) bilateral temporoparietal junction (neurosynth term: theory of mind), with peak voxels in left (-60,-56,14) and right TPJ (56,-58,20) and sizes of 1031 and 1416 voxels, respectively. It is noteworthy that the ventral part of this mask also contains voxels from posterior superior temporal sulcus (STS). For simplicity, we refer to this mask nevertheless as "TPJ". (2) dorsomedial PFC (neurosynth term: theory of mind), with a peak voxel in medial DMPFC (-2, 28, 62) and a size of 3175 voxels. The ROI mask for the amygdala, which is an anatomically well-defined region, was created via: bilateral amygdala (AAL) with sizes of 439 (left) and 492 (right) voxels. Additional exploratory analyses were conducted in regions involved in evaluating the anticipated outcomes of choice options, such as ventral striatum and vmPFC [66] (neurosynth term: "reward") using the following masks: bilateral ventral striatum (combined mask of AAL putamen and caudate up to z = 8), with sizes of 3239 (left) and 3429 (right) voxels, respectively; ventromedial PFC (neurosynth search term: ventromedial) with a peak voxel in medial vmPFC (8, 24, -12), with a size of 1327 voxels.

Furthermore, to identify whether extended networks outside our regions of interest show effects of interest, we conducted whole brain analyses at an FWE-corrected extent threshold of $p < 0.05$ ($k > 489$, initial cluster-forming height threshold $p < 0.005$). Finally, to characterize activation patterns of interest, such as time courses and activation differences due to aversive emotion, regression coefficients (beta weights) for the canonical HRF regressors were extracted with rfxplot [67] from 6 mm spheres

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

around individual subjects' peak voxel that showed significant effects of interest on BOLD responses and functional connectivity. Follow-up tests that characterize the single components of significant interaction effects were conducted in neuroimaging space via tests of simple effects of interest.

PPI analyses

Psychophysiological Interaction analyses were conducted using the generalized form of context-dependent psychophysiological interactions toolbox (gPPI toolbox [44]), using the same statistical model as outlined above. All voxels that survived SV FWE-correction for the interaction contrast in left TPJ (-60, -54, 19, k=95) were used as seed region (shown in blue color in Figure 2c). To obtain an estimate of neural activity within the seed region, the BOLD signal from the seed region was extracted, corrected by removing effects of noise covariates, and deconvolved [68]. Psychological interaction regressors for each of the task type and stimulation intensity combinations [control decisions during (1) weak and (2) strong stimulation, trust decisions during (3) weak and (4) strong stimulation] were created by multiplying the estimated neural activity during the relevant decisions with condition-specific on- and offset times convolved with the canonical HRF. A new GLM was then estimated for each subject that consisted of the original design matrix with the addition of the four psychological interaction regressors and the time course from the seed region.

To investigate the impact of aversive emotion on trust-specific functional connectivity of the left TPJ, we probed the functional connectivity data for an interaction between threat and game type. To investigate the interaction between threat and game type, we entered the contrast estimates obtained from first level PPI models into a flexible factorial model with the factors game type (trust, non-social control), threat (absent, present), and separate covariates reflecting mean transfer in each condition. A subject factor was also included in the model. Given that we were particularly interested in trust-specific changes in functional connectivity, we first contrasted the covariates reflecting mean transfers in the trust game and mean transfers in the non-social control task in the absence of threat ($\text{Trust}_{\text{no threat}}$ > NS $\text{Control}_{\text{no threat}}$). This comparison identifies regions for which connectivity with the TPJ correlates more strongly with mean transfers in the trust game than with mean transfers in the non-

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

social control game. As the next step, we then examined how threat of shock changed the relationship between TPJ connectivity and mean transfers in the trust game, by examining the interaction between game type and threat estimated over the covariates. We illustrate these results in Figure 3, by regression plots generated with coefficients reflecting functional connectivity strength for each of the conditions, extracted from 6-mm spheres around the peak voxel of the interaction contrast. The displayed regression plots were generated by the following regression model implemented in R (using the Regression Modeling Strategies package, RMS):

$$y_{ik} = \beta_0 + \beta_1\ Transfer_{ik} + \beta_2\ GameType_k + \beta_3\ Threat_k + \beta_4\ (Transfer_{ik} * GameType_k) + \beta_5 (Transfer_{ik} * Threat_k) + \beta_6 (GameType_k * Threat_k) + \beta_7 (Transfer_{ik} * GameType_k * Threat_k) + \varepsilon_{ik}$$

The dependent variable y_{ik} is the functional connectivity strength between a given brain region and the TPJ for individual *i* in Game type k. Transfer_{ik} is the mean amount sent by individual *i* in Game type k. Game type is a dummy variable encoding whether decisions were made in the trust or the control task (1 indicates trust, 0 indicates non-social control task). Threat is a dummy variable encoding whether decisions were made in the presence or absence of threat (1 indicates presence, 0 indicates absence of threat). In this regression, the coefficient for Transfer_{ik} (β_1) measures the slope of the relationship between TPJ connectivity and mean transfers in the absence of threat in the control task (see blue lines in Figures 3a – 3c), and the sum of the coefficients for Transfer_{ik} (β_1) and the interaction term between Transfer_{ik} * GameType_k (β_4) measures the trust-related slope increase of the relationship between TPJ connectivity and mean transfers in the absence of threat (see orange lines in Figures 3a – 3c). Equivalent analyses were performed to probe for significant differences between the threat and the no-threat condition in the trust game in the relationship between functional TPJ connectivity and mean transfer levels. Here, the sum of the coefficients for Transfer_{ik} (β_1), the interaction term between Transfer_{ik} * GameType_k (β_4), the interaction term between Transfer_{ik} * $Threat_k$ (β_5), and the interaction term between Transfer_{ik} * GameType_k * Threat_k (β_7) measures the slope of the relationship between TPJ connectivity and mean trust in the presence of threat (see red line in Figures 3d).

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. 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,** 1611–1618 (2005).
2. Bohnet, I. & Zeckhauser, R. Trust, risk and betrayal. *Journal of Economic Behavior & Organization* **55,** 467–484 (2004).
3. Aimone, J. A., Houser, D. & Weber, B. Neural signatures of betrayal aversion: an fMRI study of trust. *Proceedings of the Royal Society B: Biological Sciences* **281,** 20132127–20132127 (2014).
4. Fehr, E. On the Economics and Biology of Trust. *Journal of the European Economic Association* **7,** 235–266 (2009).
5. King-Casas, B. *et al.* The rupture and repair of cooperation in borderline personality disorder. *Science* **321,** 806–810 (2008).
6. Gromann, P. M. *et al.* Trust versus paranoia: abnormal response to social reward in psychotic illness. *Brain* **136,** 1968–1975 (2013).
7. FeldmanHall, O., Raio, C. M., Kubota, J. T., Seiler, M. G. & Phelps, E. A. The Effects of Social Context and Acute Stress on Decision Making Under Uncertainty. *Psychological Science* **26,** 1918–1926 (2015).
8. Mislin, A., Williams, L. V. & Shaughnessy, B. A. Motivating trust: Can mood and incentives increase interpersonal trust? *Journal of Behavioral and Experimental Economics* **58,** 11–19 (2015).
9. Loewenstein, G. Emotions in Economic Theory and Economic Behavior. **90,** 426–432 (2000).
10. Kuhnen, C. M. & Knutson, B. The Influence of Affect on Beliefs, Preferences, and Financial Decisions. **46,** 605–626 (2011).
11. Schulreich, S. *et al.* Music-evoked incidental happiness modulates probability weighting during risky lottery choices. *Front. Psychol.* **4,** 981 (2014).
12. Phelps, E. A. in *Neuroeconomics* (eds. Glimcher, P. W., Fehr, E., Camerer, C. F. & Poldrack, R. A.) 233–250 (Elsevier, 2008).
13. Knutson, B., Wimmer, G. E., Kuhnen, C. M. & Winkielman, P. Nucleus accumbens activation mediates the influence of reward cues on financial risk taking. *NeuroReport* **19,** 509–513 (2008).
14. Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E. & Barrett, L. F. The brain basis of emotion: A meta-analytic review. *Behav Brain Sci* **35,** 121–143 (2012).
15. Pessoa, L. & Adolphs, R. Emotion processing and the amygdala: from a 'low road' to ‘many roads’ of evaluating biological significance. *Nat Rev Neurosci* **11,** 773–783 (2010).
16. Kable, J. W. & Glimcher, P. W. The Neurobiology of Decision: Consensus and Controversy. *Neuron* **63,** 733–745 (2009).
17. Rangel, A. & Hare, T. Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology* **20,** 262–270 (2010).
18. Berg, J., Dickhaut, J. & McCabe, K. Trust, Reciprocity, and Social History. *Games and Economic Behavior* **10,** 122–142 (1995).
19. Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U. & Fehr, E. Oxytocin increases trust in humans. *Nature* **435,** 673–676 (2005).
20. Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U. & Fehr, E. Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* **58,** 639–650 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

21. Bogdan, R. & Pizzagalli, D. A. Acute stress reduces reward responsiveness: implications for depression. *BPS* **60,** 1147–1154 (2006).
22. Grillon, C., Ameli, R., Foot, M. & Davis, M. Fear-potentiated startle: Relationship to the level of state/trait anxiety in healthy subjects. *Biological Psychiatry* **33,** 566–574 (1993).
23. Schmitz, A. & Grillon, C. Assessing fear and anxiety in humans using the threat of predictable and unpredictable aversive events (the NPU-threat test). *Nat Protoc* **7,** 527–532 (2012).
24. Westermann, R., Spies, K., Stahl, G. & Hesse, F. W. Relative effectiveness and validity of mood induction procedures: a meta□analysis. *Eur. J. Soc. Psychol.* **26,** 557–580 (1996).
25. Martin, M. On the induction of mood. *Clinical Psychology Review* **10,** 669–697 (1990).
26. Fehr, E. & Camerer, C. F. Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences* **11,** 419–427 (2007).
27. Sripada, C. S. *et al.* Functional neuroimaging of mentalizing during the trust game in social anxiety disorder. *NeuroReport* **20,** 984–989 (2009).
28. Stanley, D. A. *et al.* Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367,** 744–753 (2012).
29. McCabe, K., Houser, D., Ryan, L., Smith, V. & Trouard, T. A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences* **98,** 11832–11835 (2001).
30. Krueger, F., Grafman, J. & McCabe, K. Neural correlates of economic game playing. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **363,** 3859–3874 (2008).
31. Saxe, R. & Kanwisher, N. People thinking about thinking peopleThe role of the temporo-parietal junction in 'theory of mind'. *NeuroImage* **19,** 1835–1842 (2003).
32. Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A. & Saxe, R. Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences* **107,** 6753–6758 (2010).
33. Mitchell, J. P., Macrae, C. N. & Banaji, M. R. Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron* **50,** 655–663 (2006).
34. Gromann, P. M. *et al.* Trust versus paranoia: abnormal response to social reward in psychotic illness. *Brain* **136,** 1968–1975 (2013).
35. Stanley, D. A. *et al.* Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367,** 744–753 (2012).
36. Ochsner, K. N., Silvers, J. A. & Buhle, J. T. Functional imaging studies of emotion regulation: a synthetic review and evolving model of the cognitive control of emotion. *Annals of the New York Academy of Sciences* **1251,** E1–E24 (2012).
37. Pessoa, L. Embedding reward signals into perception and cognition. **4,** 1–8 (2010).
38. LeDoux, J. The amygdala. *Current Biology* **17,** R868–74 (2007).
39. Engell, A. D., Haxby, J. V. & Todorov, A. Implicit Trustworthiness Decisions:

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Automatic Coding of Face Properties in the Human Amygdala. *Journal of Cognitive Neuroscience* **19,** 1508–1519 (2007).

40. Winston, J. S., Strange, B. A., O'Doherty, J. & Dolan, R. J. Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nat Neurosci* **5,** 277–283 (2002).
41. Behrens, T. E. J., Hunt, L. T. & Rushworth, M. F. S. The Computation of Social Behavior. *Science* **324,** 1160–1164 (2009).
42. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods* **8,** 665–670 (2011).
43. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage* **76,** 412–427 (2013).
44. McLaren, D. G., Ries, M. L., Xu, G. & Johnson, S. C. A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches. *NeuroImage* **61,** 1277–1286 (2012).
45. Van Overwalle, F. Social cognition and the brain: a meta-analysis. *Hum. Brain Mapp.* **30,** 829–858 (2009).
46. Bohnet, I. & Zeckhauser, R. Trust, risk and betrayal. *Journal of Economic Behavior & Organization* **55,** 467–484 (2004).
47. Bohnet, I., Greig, F., Herrmann, B. & Zeckhauser, R. Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review* **98,** 294–310 (2008).
48. Krueger, F., Grafman, J. & McCabe, K. Neural correlates of economic game playing. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363,** 3859–3874 (2008).
49. Fett, A.-K. J., Gromann, P. M., Giampietro, V., Shergill, S. S. & Krabbendam, L. Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Social Cognitive and Affective Neuroscience* **9,** 395–402 (2014).
50. McCabe, K. A., Rigdon, M. L. & Smith, V. L. Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization* **52,** 267–275 (2003).
51. Behrens, T. E. J., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. S. Associative learning of social value. *Nature* **456,** 245–249 (2008).
52. Carter, R. M., Bowling, D. L., Reeck, C. & Huettel, S. A. A Distinct Role of the Temporal-Parietal Junction in Predicting Socially Guided Decisions. *Science* **337,** 109–111 (2012).
53. Morishima, Y. *et al.* Linking Brain Structure and Activationin Temporoparietal Junctionto Explain the Neurobiology of Human Altruism. *Neuron* **75,** 73–79 (2012).
54. Pessoa, L. Emotion and cognition and the amygdala: From “what is it?” to “what's to be done?”. *Neuropsychologia* **48,** 3416–3429 (2010).
55. Sander, D., Grafman, J. & Zalla, T. The Human Amygdala: An Evolved System for Relevance Detection : Reviews in the Neurosciences. *Reviews in the Neurosciences* 303–316 (2003). doi:10.1515/REVNEURO.2003.14.4.303","pf:authenticationStatus":"logged-in","pf:selectedLanguage":"English","pf:contentCategory":"nlm-article"}
56. Lerner, J. S., Li, Y., Valdesolo, P. & Kassam, K. S. Emotion and decision making. *Annu. Rev. Psychol.* **66,** 799–823 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

57. Phelps, E. A., Lempert, K. M. & Sokol-Hessner, P. Emotion and decision making: multiple modulatory neural circuits. *Annu. Rev. Neurosci.* **37,** 263–287 (2014).
58. Yarkoni, T. & Braver, T. S. in *The Handbook of Individual Differences in Cognition* (eds. Gruszka, A., Matthews, G. & Szymura, B.) 87–108 (2010).
59. Brooks, A. M. *et al.* From Bad to Worse: Striatal Coding of the Relative Value of Painful Decisions. *Front. Neurosci.* **4,** (2010).
60. Singer, T. *et al.* Empathic neural responses are modulated by the perceived fairness of others. *Nature* **439,** 466–469 (2006).
61. Fareri, D. S., Chang, L. J. & Delgado, M. R. Computational Substrates of Social Value in Interpersonal Collaboration. *Journal of Neuroscience* **35,** 8170–8180 (2015).
62. Pruessmann, K. P., Weiger, M., Scheidegger, M. B. & Boesiger, P. SENSE: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine* **42,** 952–962 (1999).
63. Henson, R. N. A., Price, C. J., Rugg, M. D., Turner, R. & Friston, K. J. Detecting latency differences in event-related BOLD responses: application to words versus nonwords and initial versus repeated face presentations. *NeuroImage* **15,** 83–97 (2002).
64. Rilling, J. K. & Sanfey, A. G. The Neuroscience of Social Decision-Making. *Annu. Rev. Psychol.* **62,** 23–48 (2011).
65. Carter, R. M. & Huettel, S. A. A nexus model of the temporal– parietal junction. *Trends in Cognitive Sciences* **17,** 328–336 (2013).
66. Levy, D. J. & Glimcher, P. W. The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology* 1–12 (2012). doi:10.1016/j.conb.2012.06.001
67. Gläscher, J. Visualization of Group Inference Data in Functional Neuroimaging. *Neuroinform* **7,** 73–82 (2009).
68. Gitelman, D. R., Penny, W. D., Ashburner, J. & Friston, K. J. Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. *NeuroImage* **19,** 200–207 (2003).

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

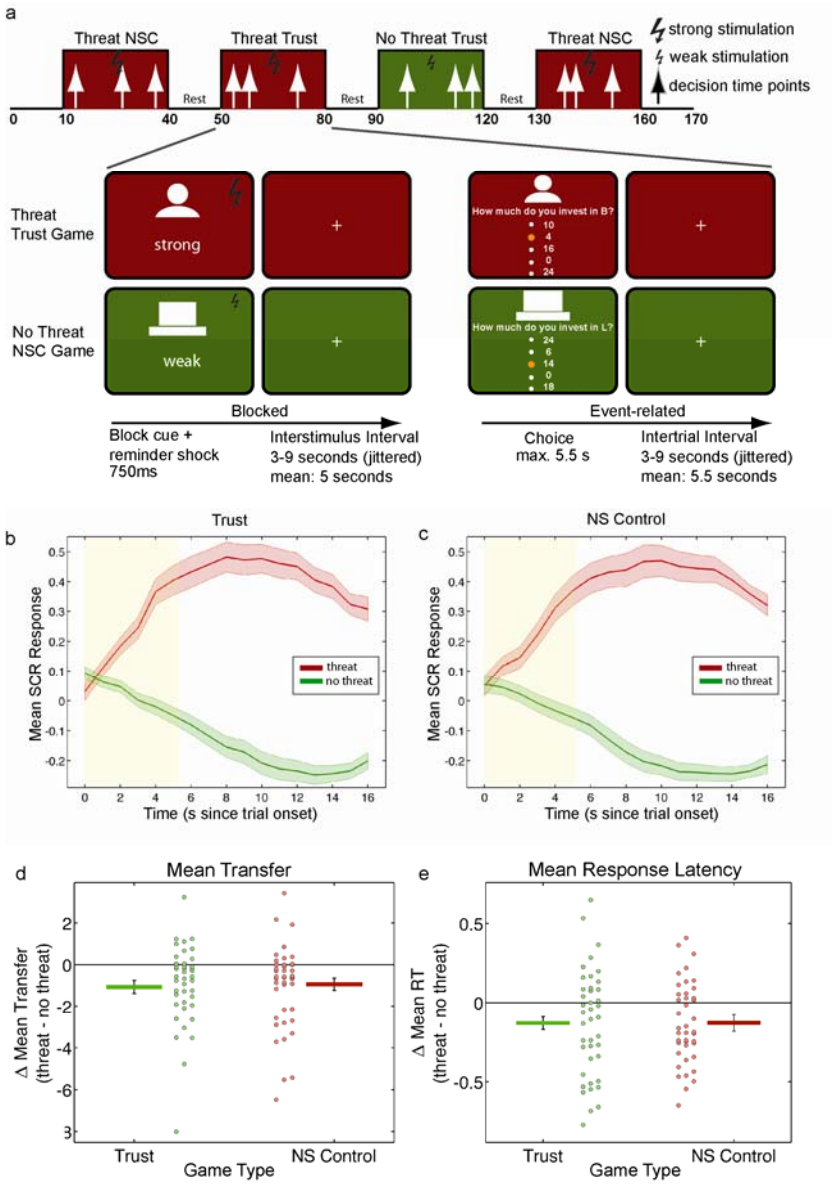

Figure 1. Experimental task, electrophysiological and behavioral findings. (a) Schematic representation of hybrid fMRI design, trial sequence and timing (see Methods). Subjects faced blocks of trust (human icon) and non-social control (NSC, computer icon) trials in random order. During trust and NSC blocks subjects expected either strong ("threat") or weak ("no-threat") tactile stimulation at unpredictable times. At the beginning of each block, a 750-ms visual cue followed by tactile stimulation reminded subjects of the game type (trust or non-social control) and stimulation intensity (weak or strong) for the current block. On each trial, subjects chose how much of their endowment of 24 CHF to transfer to a stranger (trust game), or invest in an ambiguous lottery that provided a 40-60% probability of returning an amount greater than the investment (NSC game). (b-c) The threat of an aversive tactile stimulation, not the shock itself (see Methods), leads to a strong increase in skin conductance responses (SCR) in (b) the trust game ($p < 0.0001$) and (c) the non-social control game ($p < 0.0001$), (d) In the threat condition (relative to the no-threat condition) subjects transferred significantly less to an anonymous stranger in the trust game ($p < 0.005$, reduction due to threat in 71% of subjects) and invested less into an ambiguous lottery in the non-social control game ($p < 0.005$, reduction due to threat in 73% of subjects). These results are driven by the emotional arousal induced by the threat of a shock and not by the actual experience of shocks shortly before choice (Table S1). (e) In the threat condition (relative to the no-threat condition) subjects made their decisions significantly faster in both the trust ($p < 0.005$) and the control ($p < 0.05$) game. Dot plots reflect the difference between mean transfer (d) and response latency (e) in the threat compared to the no threat condition for the same subject to illustrate the *reduction* of mean transfer and response latency due to threat.

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

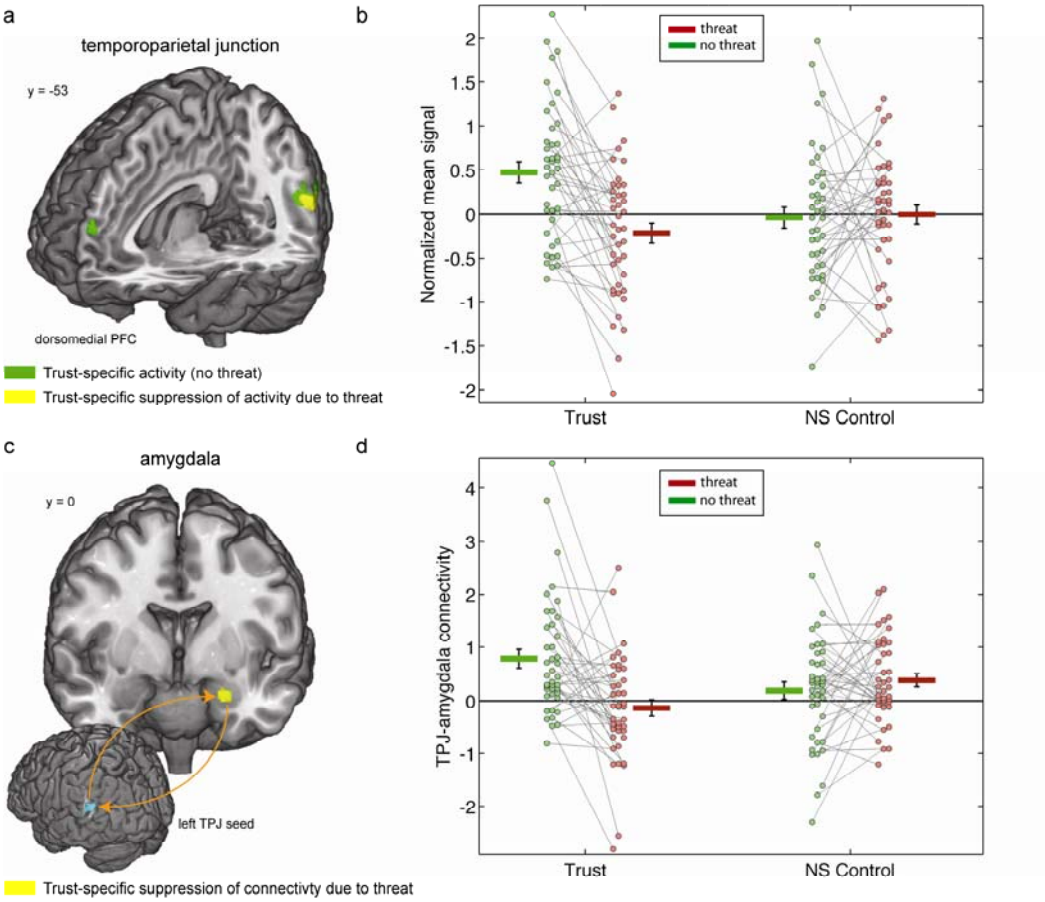

Figure 2. The impact of aversive emotion on trust-specific TPJ activity and connectivity. Panel (a) depicts brain regions (in green) that are selectively involved in trust compared to the non-social control task (see also Table S2). These regions comprise the left temporoparietal junction (peak at xyz = -57, -60, 27), the dorsomedial PFC (peak at xyz = -9, 60, 18), and the ventromedial PFC (peak at xyz = 9, 32, -12, not shown here). Importantly, aversive emotions induced by the threat of a shock reduced activation in left TPJ (relative to "no threat") significantly more during the trust game than in the nonsocial control game (significant interaction effect, peak at xyz = -60, -54, 19). Voxels whose activity reflects this interaction effect are shown in yellow. All regions are depicted at $p<0.05$ SVC-FWE-corrected (see methods). Threat-induced reduction of TPJ activity was observed in 78% of subjects during trust decisions (downward-sloping connecting lines), and in 44% of subjects during non-social control decisions, as shown in panel (b). The parameter estimates in (b) are extracted from a sphere (6-mm radius) around individual peaks within the TPJ cluster marked in yellow in panel a. (c) The left amygdala (peak at xyz = -22, -9, -15, see Table S4) shows significantly stronger connectivity with TPJ during trust relative to control when aversive emotion is absent. This coupling is disrupted by the threat of a shock specifically during trust as compared to the non-social control task (significant interaction effect; peak at xyz = -26, 0, -23). All regions are depicted at $p<0.05$ SVC-FWE-corrected (see methods). Threat-induced reduction of TPJ-amygdala connectivity was observed in 76% of subjects during trust decisions (downward-sloping connecting lines), and in 44% of subjects during non-social control decisions, as shown in panel (d). The parameter estimates are extracted from a 6-mm sphere around the individual peaks within the amygdala cluster marked in yellow in panel c, to visualize the specific effects of aversive emotion on functional connectivity between the left TPJ and left amygdala during decisions in the trust game. Dot plots in panels (e) and (d) reflect individual subject mean activation in each condition and are connected to illustrate the suppression of activity due to threat for each subject.

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

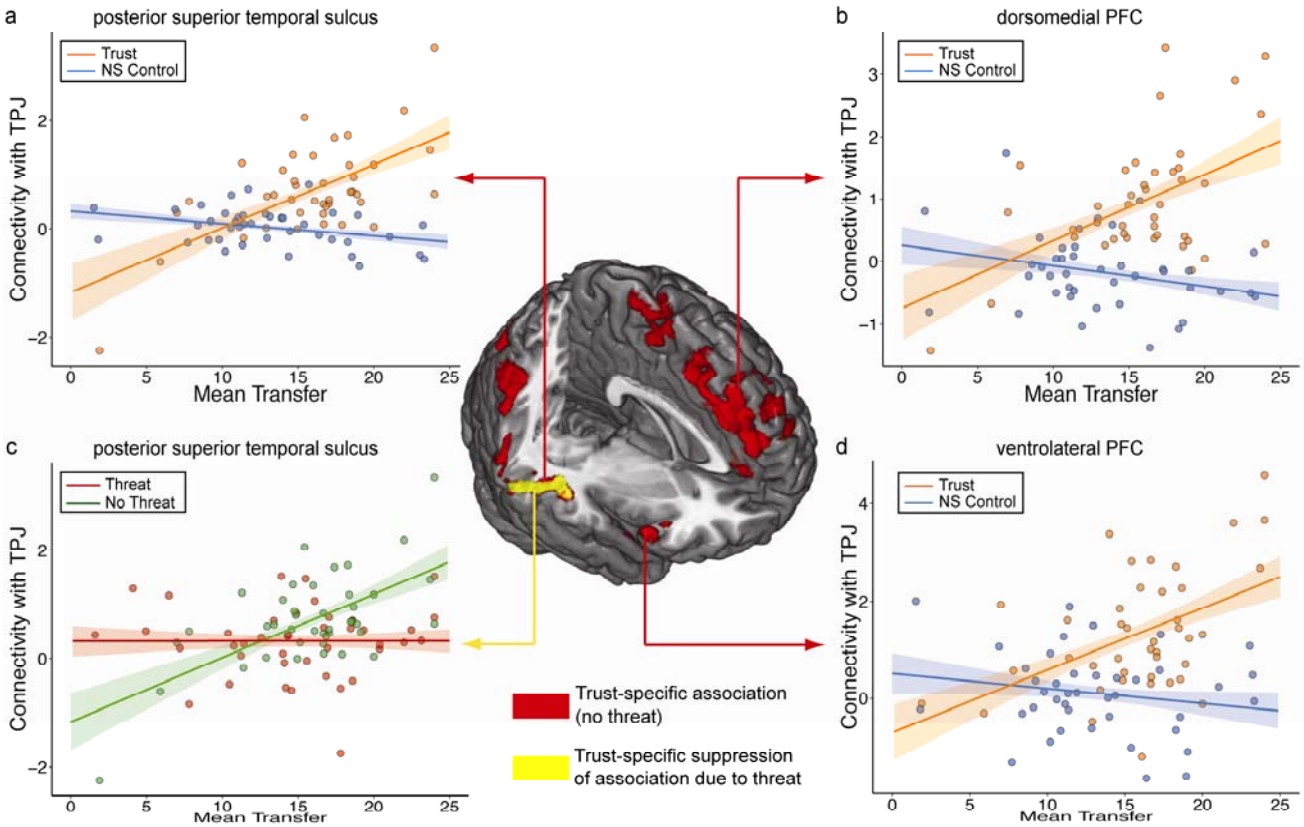

Figure 3. Trust-specific functional connectivity (a-b, d) and threat-induced breakdown of connectivity (c) between TPJ and a network of target regions. (a-b, d) Results from the simple effects contrast of trust vs NS control in the absence of threat are shown in red activation clusters: Connectivity between left TPJ and its targets is positively associated with trust taking (orange regression lines) *during the no-threat condition* in (a) posterior superior temporal sulcus (peak at xyz = 63, -45,6), (b) dorsomedial PFC (peak at xyz = -3, 18, 55), (d) bilateral ventrolateral PFC (left peak at xyz = -50, 23, -8; right peak at xyz = 57, 20, 10). In contrast, mean transfers (i.e. investments) during the non-social (NS) control task (blue regression lines) are associated with reduced connectivity strength between TPJ and these regions. In all cases, the correlation between mean transfer and connectivity strength is stronger in the trust game compared to the non-social control task (whole brain analysis, $p < 0.05$, FWE corrected at cluster level, see Table S6a). The intraparietal sulcus (peak at xyz = 46, -58, 45) and dorsolateral PFC (peak at xyz = 50, 12, 37) show a similar pattern, but are not shown in the figure. (c) The results from the interaction contrast reflecting a trust-specific breakdown of the association between mean trust and TPJ connectivity is shown in the yellow activation cluster in STS: Aversive emotion causes a breakdown of the association between TPJ-pSTS connectivity and mean trust (shown in yellow). The correlation between mean trust levels and TPJ-pSTS connectivity is stronger in the no threat compared to the threat condition (peak at xyz = 64, -43, 4; whole brain analysis, $p < 0.05$, FWE corrected at cluster level, see Table S6b). Specifically, there is a positive association between TPJ-pSTS connectivity and the mean trust level when distortionary aversive emotion is absent (green regression line), which is eliminated by threat (red regression line). This suggests that connectivity between TPJ and its target region in pSTS supports general trust taking only in the absence of threat. The regression lines in (a-d) predict functional connectivity strength as a function of mean transfer levels based on an extended OLS model that estimates both the *slope* of the relationship between mean transfers and functional connectivity in the non-social control task and the *increase* in this relationship in the trust task (relative to the non-social control task). For this purpose we extracted the data from 6 mm spheres around individual interaction peak voxels (see online methods). Confidence bounds around regression lines reflect 95% confidence intervals around the model fit.

bioRxiv preprint doi: https://doi.org/10.1101/129130; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

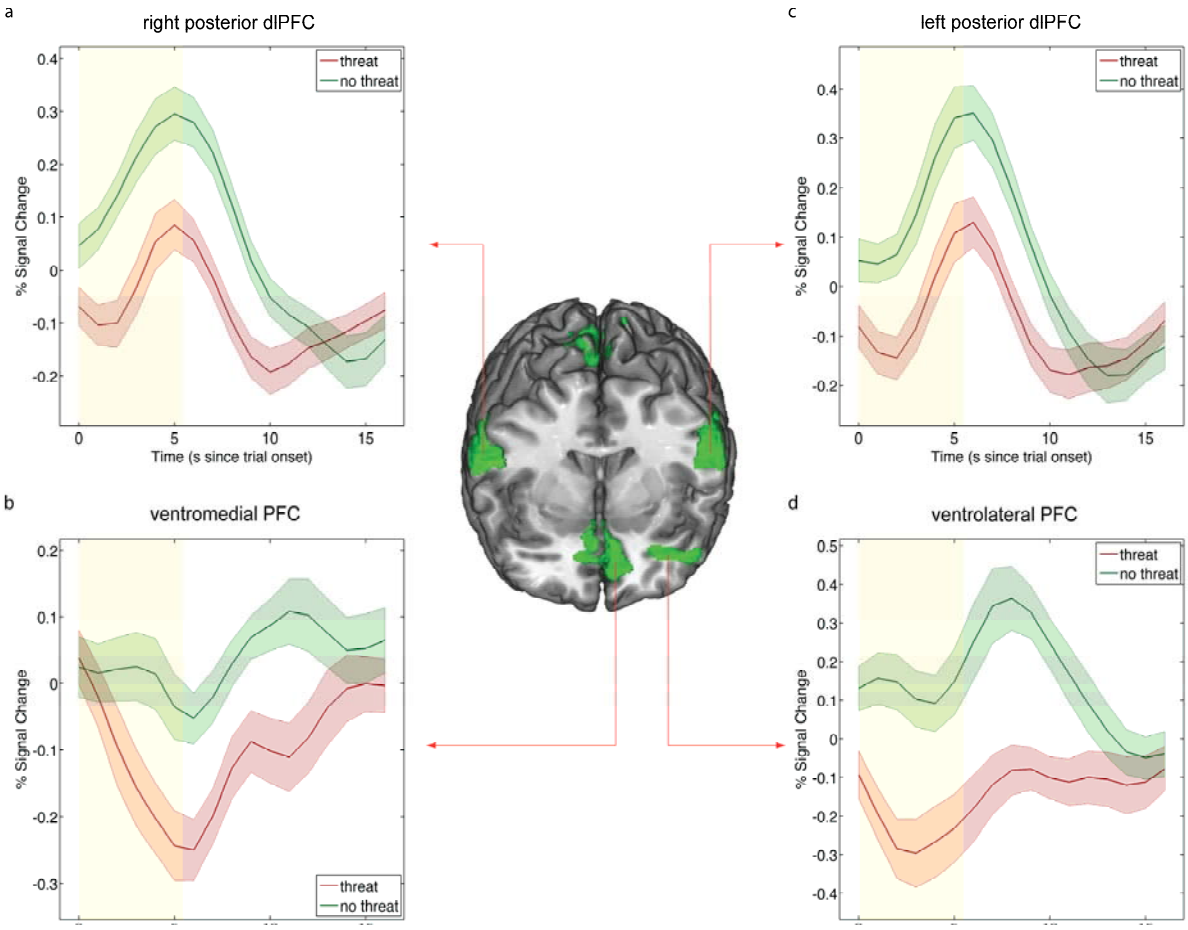

Figure 4. The impact of aversive emotion on choice-domain independent neural correlates of decision-making. We tested the main effect of aversive emotion on the neural correlates of decision-making independent of the choice domain (social and non-social). This analysis revealed a domain-general network consisting of bilateral posterior dlPFC (panel a, right peak at xyz = -62, -4, 18, and (panel c, left peak at xyz = 62, -6, 28), and a large cluster in ventral anterior prefrontal cortex that includes vmPFC (panel b, peak at xyz = 10, 44, -8) and left vlPFC (panel d, peak at xyz = -48, 41, -8). These regions show significant threat-related suppression (no threat > threat, regions shown in green) in choice-related activity during both trust and non-social control trials. Additional regions that are shown in Supplemantary Figure 4 include left amygdala (panel c, peak at xyz = -24, -15, -23), posterior paracentral lobule (peak at xyz = 4 -36, 69). Time courses reflect choice-domain independent activity that shows suppressions due to the aversive emotion during decisions in both trust and non-social control trials. To illustrate the equivalent effect of aversive emotion, Supplementary Figures 4 and 5 show activity for both trust and control trials in separate graphs. Time courses were extracted from 6 mm spheres around peak voxels. The 5.5-second choice period is displayed in yellow.