NEUROSCIENCE

The neural circuitry of affect-induced distortions of trust

Jan B. Engelmann[1,2]*, Friederike Meyer[2], Christian C. Ruff[2†], Ernst Fehr[2*†]

Aversive affect is likely a key source of irrational human decision-making, but still, little is known about the neural circuitry underlying emotion-cognition interactions during social behavior. We induced incidental aversive affect via prolonged periods of threat of shock, while 41 healthy participants made investment decisions concerning another person or a lottery. Negative affect reduced trust, suppressed trust-specific activity in the left temporo-parietal junction (TPJ), and reduced functional connectivity between the TPJ and emotion-related regions such as the amygdala. The posterior superior temporal sulcus (pSTS) seems to play a key role in mediating the impact of affect on behavior: Functional connectivity of this brain area with left TPJ was associated with trust in the absence of negative affect, but aversive affect disrupted this association between TPJ-pSTS connectivity and behavioral trust. Our findings may be useful for a better understanding of the neural circuitry of affective distortions in healthy and pathological populations.

INTRODUCTION

Trust pervades almost every aspect of human social life. It plays a decisive role in families, organizations, markets, and the political sphere. Without trust, families fall apart, organizations are inefficient, market transactions are costly, and political leaders lack public support. 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 and its distortions in psychiatric disorders [e.g., (*4*, *5*)], there are still large gaps in our knowledge about the impact of our emotions on trust and particularly the underlying neural circuitry.

We focus here on incidental affect, which has been shown to distort choice behavior for financial and other types of social decisions (*6*, *7*). Incidental emotions are of particular interest due to 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. Prominent theoretical accounts (*8*–*10*) distinguish this incidental affect from anticipatory affect that reflects how decision-makers expect to feel about the outcomes of their decisions. While recent research has made much progress in outlining the neural underpinnings of affective processes on the one hand (*11*) and of decision-making on the other (*12*), the neural interactions between affective and cognitive processes that support choice have largely been explored from theoretical perspectives (*13*, *14*). It is therefore important to directly examine the behavioral and neural mechanisms by which incidental affect distorts decisions to trust.

To study the behavioral impact and the underlying neural circuitry of affect-induced distortions of trust, we used a modified version of the well-established trust game that has also been used in a number of previous imaging studies (*1*, *15*, *16*). 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 among the two players. In the second stage, the trustee is informed about the total amount that he received and then needs to decide how much of this money to send back (nothing, parts of it, or all of it). Thus, while the investor's transfer increases the total amount of money available to both parties, the investor also faces the risk of not benefitting at all 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.

Trust decisions involve both a financial risk due to the possibility of losing the invested money and a social risk of being betrayed by an untrustworthy trustee (*2*, *3*). The latter social risk is specific to trust, as compared to nonsocial (NS) types of risky choices. To enable clear identification of the impact of incidental emotion on the mechanisms specifically involved in trust, it is important to include a well-matched NS control task that has equivalent financial payouts without the social risk of betrayal. For this reason, our participants also faced an NS 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" (*16*). This back transfer was determined using an algorithm that drew random samples from the probability distribution of the trustees' decisions in the trust condition (see Materials and Methods). Thus, the choice options and the profitability of the investor's transfer in the NS control and trust condition were exactly the same, and the distinguishing feature between the trust and the NS control game was the unique possibility of betrayal by the interaction partner in the trust game (*2*). This difference between the trust and the NS 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 nonhuman symbol (in the NS control game).

The unique possibility of being betrayed by the trustee provides a strong incentive to avoid this betrayal (*2*). Therefore, the investor needs to assess the likelihood of this betrayal when deciding how

[1]Center for Research in Experimental Economics and Political Decision Making (CREED), Amsterdam School of Economics and Amsterdam Brain and Cognition (ABC), University of Amsterdam and the Tinbergen Institute, Amsterdam, Netherlands. [2]Laboratory for Social and Neural Systems Research, Department of Economics, University of Zürich, Zürich, Switzerland.
*Corresponding author. Email: ernst.fehr@econ.uzh.ch (E.F.); jbengelmann@gmail.com (J.B.E.)
†These authors share senior authorship.

much money to entrust to the interaction partner. This assessment of betrayal likelihood is accomplished by mentalizing, i.e., taking the perspective of the trustee to estimate how she will react to given transfers. In contrast, no such processes are necessary to make a decision in the NS control task. Recent neuroimaging studies (*5*, *17*, *18*) have confirmed that decisions to trust may require mentalizing, since brain areas commonly found to be involved in mentalizing [including the dorsomedial prefrontal cortex (dmPFC), superior temporal sulcus (STS), and temporoparietal junction (TPJ)] are activated during the trust game. We therefore expected that aversive affect may exert its effect on decisions to trust by affecting trust-specific activation in regions involved in representing other people's mental states, including the TPJ and dmPFC (*19*, *20*).

To investigate the impact of aversive incidental affect on trust decisions, participants made decisions in either trust or NS control trials within two different affective 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 that was still noticeable, but not painful. There was thus no anticipation of aversive events in this "no threat" condition (Fig. 1A). A prolonged period of incidental aversive affect was established by administering the tactile stimulations at unpredictable time points and frequencies for 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 used in the current study has been shown to reliably induce

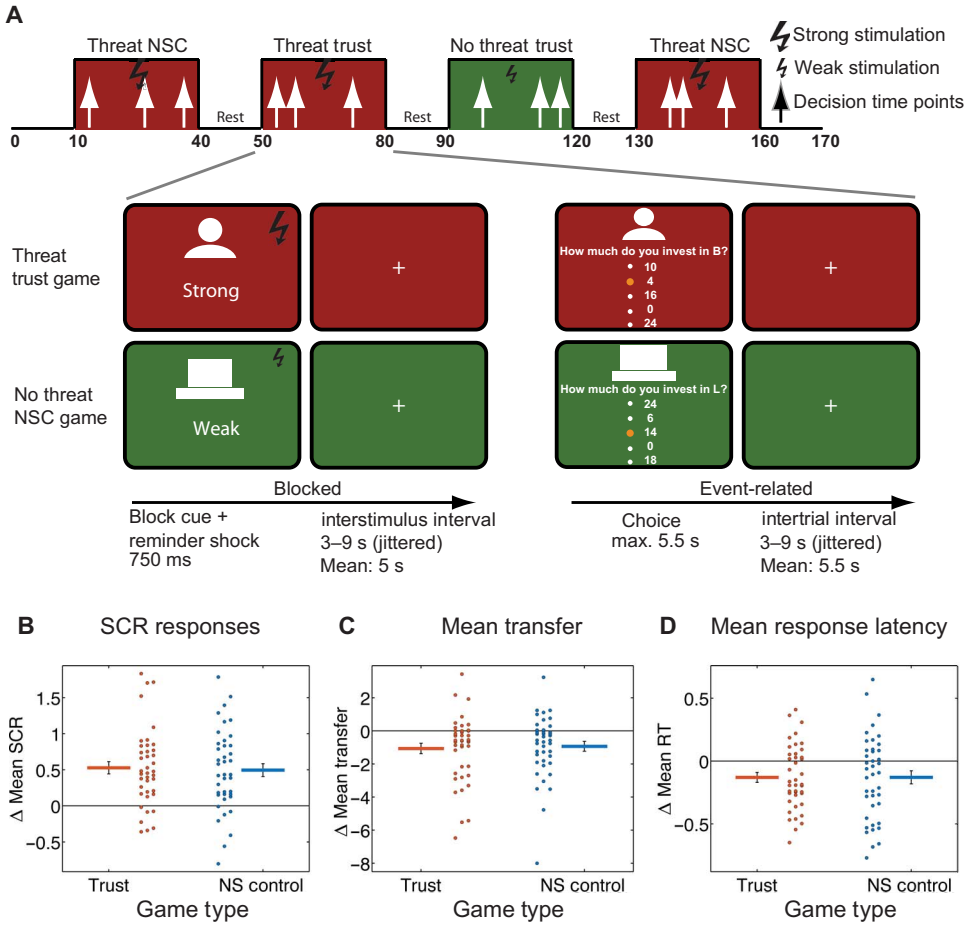

Fig. 1. Experimental task and electrophysiological and behavioral findings. (**A**) Schematic representation of hybrid functional magnetic resonance imaging (fMRI) design, trial sequence, and timing (see Materials and Methods). Participants faced blocks of trust (human icon) and NS control (NSC; computer icon) trials in random order. During trust and NS control blocks, participants 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 participants of the game type (trust or NS control) and stimulation intensity (weak or strong) for the current block. On each trial, participants 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 to 60% probability of returning an amount greater than the investment (NS control game). (**B**) The threat of an aversive tactile stimulation leads to a strong increase in skin conductance responses (SCRs) in the trust game (orange; $P < 0.0001$) and the NS control game (blue; $P < 0.0001$). (**C**) In the threat condition (relative to the no-threat condition), participants transferred significantly less to an anonymous stranger in the trust game (orange; $P < 0.005$; reduction due to threat in 71% of participants) and invested less into an ambiguous lottery in the NS control game (blue; $P < 0.005$; reduction due to threat in 73% of participants). 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). (**D**) In the threat condition (relative to the no-threat condition), participants made their decisions significantly faster in both the trust (orange; $P < 0.005$) and the control (blue; $P < 0.05$) game. Dot plots in all panels reflect the change for each participant in the presence of threat (threat–no threat) in mean SCRs (B), mean transfers (C), and mean response latencies (D). These plots thus show the enhancement of affective arousal and the reduction of mean transfer and response latency due to threat.

negative affect (*21*) and addresses the limitations of standard emotion (*22*, *23*) and stress induction (*24*, *25*) procedures as follows: First, the threat of shock provides an immediate stimulus of biological importance that triggers an aversive and automatic emotional reaction, the intensity of which can be titrated to individual subjective percepts and measured throughout the experiment using standard psychophysiological techniques. Second, unlike standard emotion and stress inductions that are typically administered only once at the start of a session, threat of shock can be turned on and off throughout the duration of the entire experiment, thereby reinstating the emotional reaction at every presentation. This disambiguates anxiety from stress recovery, which is hard to achieve with standard emotion and stress inductions (*26*). Third, threat of shock was administered within-subject in a single experimental session, therefore allowing each participant to serve as their own control. Last, tactile stimulation was administered in both the trust and the NS control condition, thus minimizing demand effects.

We conjectured that incidental aversive affect modulates trust-specific computations, particularly the simulations of the trustee's reaction to given transfers. This has potentially important implications because if incidental aversive affect disrupts the recruitment of the social cognitive processes necessary for mentalizing and perspective taking, then we are likely to observe that it also has specific effects on the neural mechanisms of trust decisions. We expected that incidental aversive affect influences trust decisions by specifically modulating neural responses in regions involved in representing other people's mental states, including the TPJ and dmPFC (*19*). Three pieces of evidence support our neuroanatomical hypothesis. First, theoretically, investors are highly motivated to estimate the likelihood of betrayal via simulations of the trustee's reaction to given transfers as a fundamental aspect of their decisions to trust. Previous research has consistently implicated TPJ and dmPFC in these simulations of others' mental states [for meta-analysis, see (*19*)]. Second, previous research has demonstrated the involvement of the same neural structures (TPJ and dmPFC) also during trust decisions and similar social interactions (*18*). Third, a conjunction of the neurosynth meta-analyses for the terms "emotion" and "theory of mind" identifies an overlap between these networks in the TPJ and dmPFC, therefore implicating these regions in both affective and social cognitive processes. Together, the above results establish these regions as prime candidates for investigations of the modulatory effects of incidental aversive affect on mentalizing during trust decisions. In addition, we explored the effects of incidental aversive affect on neural activity in the anterior insula, which has consistently been implicated in both aversive emotion (*27*) and trust decisions (*28*), and the amygdala, which has consistently (*29*) and relatively specifically (neurosynth reverse inference analysis for "aversive") been implicated in the processing of aversive emotions (*11*) and, at the same time, plays a central role in trustworthiness inferences (*30*).

RESULTS

Threat of shock induces autonomic arousal and aversive affect 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. We first ascertained that threat of shock induced emotional arousal by probing how galvanic skin conductance responses (SCRs), self-reported emotion, and brain activations changed in response to the threat of electrical stimulation (note that SCR was modeled with a general linear model that included regressors for the time points of actual shocks; see the "Skin conductance responses" section in Materials and Methods). As illustrated in Fig. 1B, mean SCRs during both trust and NS control trials were on average significantly greater during the threat condition compared to the no-threat condition (trust decisions: mean difference = 0.53, t_{39} = 6.20, $P < 0.0001$; NS control game: mean difference = 0.50, t_{39} = 5.56, $P < 0.0001$). Similar observations were made for predictable and unpredictable strong versus weak shocks (fig. S1A). Together, these results indicate significantly greater affective arousal during the threat condition relative to the no-threat condition in both social and NS game types.

The affective arousal illustrated in Fig. 1B was experienced as aversive by the participants. In an open-ended questionnaire administered after scanning, 95.12% of participants responded that they experienced aversive emotional arousal during threat blocks (fig. S1B). 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 (fig. S1C) and by the observation of enhanced SCRs following the (actual) experience of strong compared to weak tactile shocks (fig. S1A).

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

Aversive affect reduces investments during trust decisions

To identify whether the aversive affective state had a significant impact on decision-making, we first investigated mean transfer rates during trust and NS control decisions for each affective context and submitted these data to a two-way repeated-measures analysis of variance (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 and control) and threat (absent and present). Aversive affective state significantly changed transfers during both trust and NS control trials (Fig. 1D), as indicated by a significant main effect of threat ($F_{1,40} = 17.483$, $P < 0.001$, $\eta^2 = 0.304$). We also observed a significant main effect of game type ($F_{1,40} = 20.319$, $P < 0.001$, $\eta^2 = 0.337$), with larger transfers in the trust (15.37) compared to the control game (12.67). Moreover, separate pairwise comparisons (all two-tailed) showed that the threat condition led to a reduction of investments (Fig. 1D) in the trust game [$t_{40} = -3.4$, $P < 0.005$, mean transfer difference = −1.1 Swiss francs (CHF)] and in the NS 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 affect due to shock expectation, we ran several multiple regression analyses (section S1). The regression results (table S1) show that the behavioral results reported above were due to the aversive affect ($P < 0.001$) generated by the threat of shock, rather than reflecting the effect of actual shock experience immediately before decisions are taken ($P = 0.23$).

Aversive affect also led to faster decision times during both trust and NS control trials (Fig. 1E). Mean decision times were submitted to a two-way repeated-measures ANOVA (mean decision times were normally distributed as indicated by the Shapiro-Wilk normality test:

$W = 0.972$, $P = 0.402$) with the factors game type (trust and control) and threat (absent and 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 decision times in the threat relative to the no-threat condition (Fig. 1E) for both the trust game [$t_{40} = -3.3$, $P < 0.005$, mean response time (RT) difference = −0.13 s] and the NS control task ($t_{40} = -2.5$, $P < 0.05$, mean RT difference = −0.13 s).

Together, these behavioral results indicate that aversive affect significantly reduced trust, as reflected by diminished transfer rates in the trust game. In addition, aversive affect reduced transfer rates in the NS control task and decision times in both the trust and the NS 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 affect during trust and NS control trials is similar across these multiple measurement modalities, confirming that our NS condition constitutes a well-matched control for the trust game.

Aversive affect suppresses trust-related activity in TPJ

The main goal of our functional magnetic resonance imaging (fMRI) analyses was to identify the impact of aversive affect on the neural mechanisms instantiating the social cognitive processes that support trust decisions. We therefore first examined brain activation in the regions of interest (ROIs) that we conjectured (see our hypotheses in Introduction) 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 the TPJ and dmPFC (*19*). We used small-volume correction at a family-wise error (FWE)–corrected threshold of $P < 0.05$ in truly independent ROIs defined with reverse inference maps from relevant search terms on neurosynth.org (see Materials and Methods) (*31*). Where appropriate, we supplemented these hypothesis-driven ROI analyses with exploratory whole-brain analyses as outlined in Materials and Methods. Furthermore, our fMRI analyses followed these steps: (i) We tested whether our a priori ROIs are involved in the processing of social context or affect by investigating main effects; (ii) we then probed activity in these regions for the specific suppression of trust-related activity by threat by investigating interactions; (iii) interaction results were then further characterized by simple comparisons within specific contexts (e.g., threat versus no threat in the trust task). In the last part of this paper, we also examined the domain general effects (i.e., the effects that are not specific to the trust game) of aversive affect.

As a first step, our analyses confirmed that several of the conjectured regions were specifically involved in trust (versus NS control) decisions (main effect of task: $\text{trust}_{\text{no threat}} + \text{trust}_{\text{threat}} > \text{NS control}_{\text{no threat}} + \text{NS control}_{\text{threat}}$). That is, an area in the left TPJ (−57, −60, 27; $k = 9$, SV (small volume) FWE-corrected; green region in Fig. 2A and see also table S2A) showed significantly greater activation during decision-making in trust relative to control trials. We then examined which of the ROIs showed a breakdown of trust-specific activity due to aversive affect. To this end, we identified the threat-induced reduction in brain activation that was specific to the trust game with the following interaction contrast: $(\text{trust}_{\text{no threat}} > \text{trust}_{\text{threat}}) > (\text{NS control}_{\text{no threat}} > \text{NS control}_{\text{threat}})$. A region in the left TPJ showed this significant interaction effect (−60, −54, 19; $k = 34$, SV FWE-corrected; red region in Fig. 2A; see also table S2B). To further characterize this interaction, we examined post hoc the impact of aversive affect in the trust game separately from its impact in the NS control game: This revealed a suppression of activation within the left TPJ (−58, −55, 19; $k = 35$, SV FWE-corrected; table S2C) for trust decisions, but not during NS control decisions (no voxels for $\text{NSC}_{\text{no threat}} > \text{NSC}_{\text{threat}}$), even at a very liberal threshold of $P < 0.05$, uncorrected (see also fig. S2A for additional univariate analyses that underline the strength of the interaction effect in the left TPJ). These results indicate that the interaction effect is based on a selective interference of aversive affect with trust-related activity but not with activity in the NS control condition. We also found a region in the left anterior insula that showed the same interaction effect (−46, 14, −12; SV FWE-corrected; table S2B), but the only marginally significant simple effects ($\text{trust}_{\text{no threat}} > \text{trust}_{\text{threat}}$: −46, 26, 12; $k = 4$, $P = 0.078$, SV FWE-corrected; $\text{NSC}_{\text{no threat}} > \text{NSC}_{\text{threat}}$: no voxels) indicate a weaker trust-related suppression of activity compared to the TPJ.

Given that decisions involving trust rely on neural circuitry that mediates the interplay between social cognition and valuation (*1*, *32*), we also performed an exploratory analysis of the impact of aversive affect on trust-related activity within regions commonly implicated in valuation [ventromedial PFC (vmPFC) and ventral striatum] (*33*). No significant interactions were found, but tests of simple effects comparing threat and no threat during trust decisions showed reductions in trust-related activity due to aversive affect in the vmPFC and ventral striatum (table S2C).

Aversive affect suppresses trust-specific connectivity between the TPJ and amygdala

Recent studies stress the importance of the interplay between cognitive and emotional networks (*11*, *34*). Therefore, we investigated the effects of aversive affect on the connectivity between trust-relevant brain regions with psychophysiological interaction (PPI) analyses (*35*). In view of the key role of the TPJ in perspective taking and mentalizing (*19*) and the conjecture that these mental operations are important for trust 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 affect changed the functional connectivity between the TPJ and emotion processing regions, such as the amygdala. ROI analysis of a PPI analysis seeded in the TPJ confirmed that TPJ-amygdala connectivity was particularly disrupted by threat ($\text{No threat}_{\text{trust}} + \text{No threat}_{\text{NS control}} > \text{threat}_{\text{trust}} + \text{threat}_{\text{NS control}}$) in the left amygdala (−28, −6, −14; $k = 14$, $P < 0.05$, SV FWE-corrected; table S3A). Therefore, we were interested whether there was 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 ($\text{NS control}_{\text{no threat}} > \text{NS control}_{\text{threat}}$).

This analysis revealed that threat-induced aversive affect caused a connectivity change between the TPJ and a region in the left amygdala (−26, 0, −23; $k = 12$, $P < 0.05$, SV FWE-corrected; Fig. 2C shown in red) that was significantly larger in the trust game compared to the control task (see table S3B). We performed a post hoc inspection of the significant interaction in the left amygdala by investigating the effect of threat on connectivity changes for the trust and the NS control condition separately. Aversive affect disrupted functional connectivity specifically during trust (compare red versus green bars in Fig. 2D) but not during NS control decisions. A follow-up contrast investigating

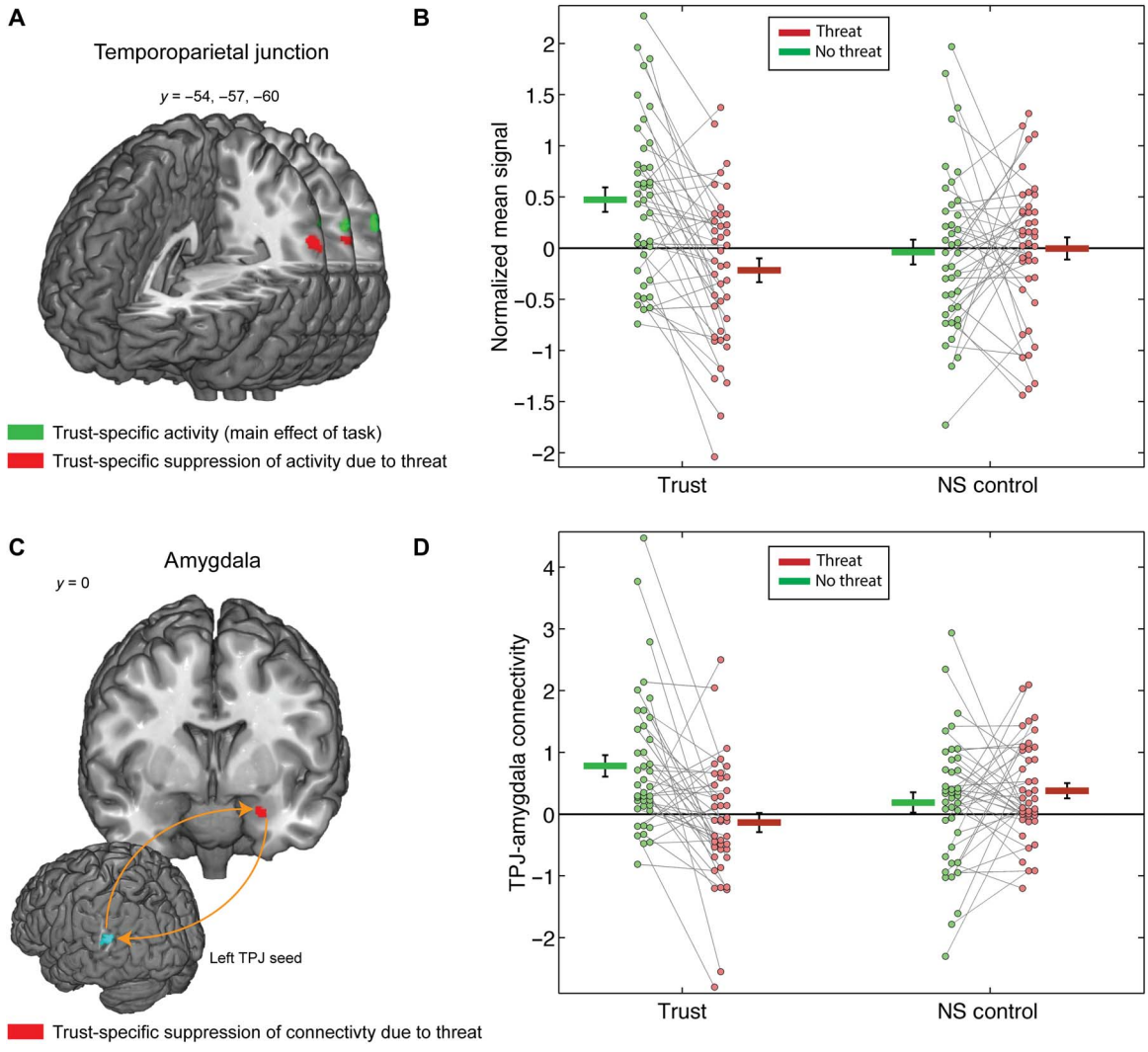

Fig. 2. The impact of aversive affect on trust-specific TPJ activity and connectivity. (**A**) The region of left TPJ (peak at *xyz* = −57, −60, 27) that is selectively involved in trust compared to the NS control task as reflected by a significant main effect of game type (shown in green; see also table S2A). Aversive affect induced by the threat of a shock reduced activation in the left TPJ (relative to no threat) significantly more during the trust game than in the NS control game (significant interaction effect; peak at *xyz* = −60, −54, 19; table S2B). Voxels whose activity reflects this interaction effect are shown in red. All regions survived SV FWE correction, $P < 0.05$ (see Materials and Methods). Threat-induced reduction of TPJ activity was observed in 78% of participants during trust decisions (downward-sloping connecting lines) and in 44% of participants during NS control decisions, as shown in (**B**). The parameter estimates in (B) are extracted from a sphere (6-mm radius) around individual peaks within the TPJ cluster marked in red in (A). (**C**) The left amygdala (peak at *xyz* = −28, −6, −14; see table S3A) shows significantly stronger connectivity with TPJ during trust relative to control decisions as reflected by a significant main effect of threat. This coupling is disrupted by the threat of a shock specifically during trust as compared to the NS control task (significant interaction effect; peak at *xyz* = −26, 0, −23; shown in red). All regions survive $P < 0.05$ SV FWE-corrected (see Materials and Methods). Threat-induced reduction of TPJ-amygdala connectivity was observed in 76% of participants during trust decisions (downward-sloping connecting lines) and in 44% of participants during NS control decisions, as shown in (**D**). The parameter estimates are extracted from a 6-mm sphere around the individual peaks within the amygdala cluster marked in yellow in (C) to visualize the specific effects of aversive affect on functional connectivity between the left TPJ and left amygdala during decisions in the trust game. Dot plots in (B) and (D) reflect individual participant mean activation in each condition and are connected to illustrate the suppression of activity due to threat for each participant.

threat effects on trust-related connectivity patterns ($trust_{no\ threat} > trust_{threat}$) confirmed the suppression of TPJ-amygdala connectivity during trust decisions (−28, −6, −14; $k = 23$, $P < 0.05$, SV FWE-corrected; table S3C). In contrast, during decisions in the NS control task, no voxels in our left amygdala ROI showed greater connectivity during no threat relative to threat, or the reverse contrast of threat versus no threat, even at a very liberal threshold of $P < 0.05$, uncorrected (see also fig. S2B 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 condition showed a significant suppression of connectivity during trust relative to NS decisions in the left amygdala (−26, 0, −23; $k = 14$, $P < 0.05$, SV FWE-corrected; table S3D), indicating that the threat-related suppression of TPJ-amygdala connectivity was also evident when comparing NS control to trust. Together, these results indicate that threat caused a specific suppression of connectivity during trust 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 occurred in the absence of suppression during NS control decisions (NS control: threat = no threat). Thus, aversive affect 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 the posterior STS, DMPFC, and ventrolateral prefrontal cortex specifically predicts behavioral trust

The above PPI analyses show the average impact of aversive affect on the functional connectivity between the TPJ and amygdala. However, as we observed strong individual differences in the functional connectivity between the TPJ and amygdala on the one hand and in mean transfer levels on the other, we 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 affect. Following our analysis approach for TPJ activity above, we first identified trust-specific brain-behavior correlations by comparing the relationship between transfer rates and TPJ functional connectivity as a function of game type, regardless of the threat condition (main effect of task: $\text{trust}_{\text{no threat}}$ + $\text{trust}_{\text{threat}}$ > $\text{NS control}_{\text{no threat}}$ + $\text{NS control}_{\text{threat}}$). Given the importance of social cognitive processes for trust decisions, we expected stronger TPJ connectivity, with other regions implicated in social cognition, such as our a priori ROIs in the dmPFC, TPJ, and amygdala, for enhanced trust. We therefore examined in our ROIs whether the relationship between mean transfers and functional TPJ connectivity is different in the trust game compared to the NS 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 found a significantly stronger positive correlation with mean transfer rates in the trust game compared to the NS control game for connectivity between the left TPJ and the right amygdala (28, 2, −20; k = 16), the dmPFC (−12, 54, 40; k = 30), the bilateral STS [right: 64, −43, 4; k = 30; left: −62, −52, −5; k = 10; note (see Materials and Methods) that the STS is the most ventral part of our TPJ mask], and the bilateral anterior insula (left: −51, 21, −6; k = 245; right: 56, 18, 1; k = 525) (all P < 0.05, SV FWE-corrected; table S4A).

For completeness, we also conducted an exploratory whole-brain analysis (FWE-corrected at the cluster level) that identified an extended network of regions (Fig. 3, A, B, and D, and table S5A), showing a difference in the relationship between individuals' mean transfers and their functional TPJ connectivity across trust and control games. This analysis extends our ROI analysis by identifying regions outside our ROIs, including the bilateral inferior frontal gyrus (IFG) (right: 42, 27, −11; k = 1422; left: −51, 21, −6; k = 343) and bilateral dorsolateral PFC (dlPFC) (right: 44, 8, 28; k = 601; left: −52, 6, 18; k = 415; see table S5A). In all these regions, we observed a positive and significantly stronger correlation between mean transfer levels and functional TPJ connectivity in the trust compared to the control task (see Fig. 3, A, B, and D, and table S5A).

Last, we also tested whether the slightly negative slopes observed in the regression lines connecting TPJ connectivity and mean transfer rates in the NS control conditions of Fig. 3 (A, B, and D) are statistically significant. To this end, we ran simple affect 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. During trust decisions in the absence of threat, on the other hand, TPJ connectivity strength with the posterior STS (pSTS) (64, −43, 4; k = 335) and dmPFC (20, 50, 39; k = 475), as well as the left IFG (−51, 27, −3; k = 429), left posterior insula (−40, 3, 7; k = 578), and right intraparietal sulcus (44, −48, 40; k = 691), positively predicted transfer rates (table S5D).

Together, the above results therefore confirm the conjecture that aversive affect suppresses trust-specific functional connectivity between the TPJ and amygdala. In addition, the larger the TPJ connectivity with key regions implicated in mentalizing (the dmPFC and right STS) and emotion (the amygdala), the more participants were willing to trust their partners on average. This association between transfer rates and TPJ connectivity was absent in the NS control game. These results thus suggest that trust involves functional communication between the TPJ and a network consisting of the amygdala, right STS, dmPFC, and bilateral ventrolateral prefrontal cortex (vlPFC).

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

How did the relationship between functional connectivity patterns in the trust network and behavioral trust change if participants were exposed to aversive affect? Our previous results showing (i) greater TPJ connectivity with dmPFC and STS not only under enhanced trust (Fig. 3) but also (ii) reduced behavioral trust under threat (Fig. 1), and (iii) suppressed trust-related TPJ activity under threat (Fig. 2) jointly lead to the prediction that aversive affect should have suppressive effects on the social cognitive machinery that supports trust decisions. We therefore expected a suppression of the relationship between behavioral trust and TPJ connectivity with other social cognition regions under conditions of threat. We tested this hypothesis in our a priori ROIs in TPJ and dmPFC. Specifically, we examined whether there was a 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 found a significant effect in the right STS [64, −43, 6; k = 7, P < 0.05, SV FWE-corrected; table S4B; recall (see Materials and Methods) that the STS is the most ventral part of our TPJ mask]. 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 affect 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 affect caused a general breakdown of the association between left TPJ connectivity and mean trust in the right pSTS (64, −43, 6; k = 31, P < 0.05, SV FWE-corrected; table S5C). In this region, therefore, there was a significantly positive relationship between left TPJ connectivity and mean trust levels during the no-threat condition (Fig. 3C, green regression line) that vanished in the presence of threat (Fig. 3C, red regression line). In contrast, during decisions in the NS control task, no voxels in any of our ROIs, as well as the STS, showed greater connectivity during no threat relative to threat, or the reverse contrast of threat versus no threat, even at a very liberal threshold of P < 0.05, uncorrected.

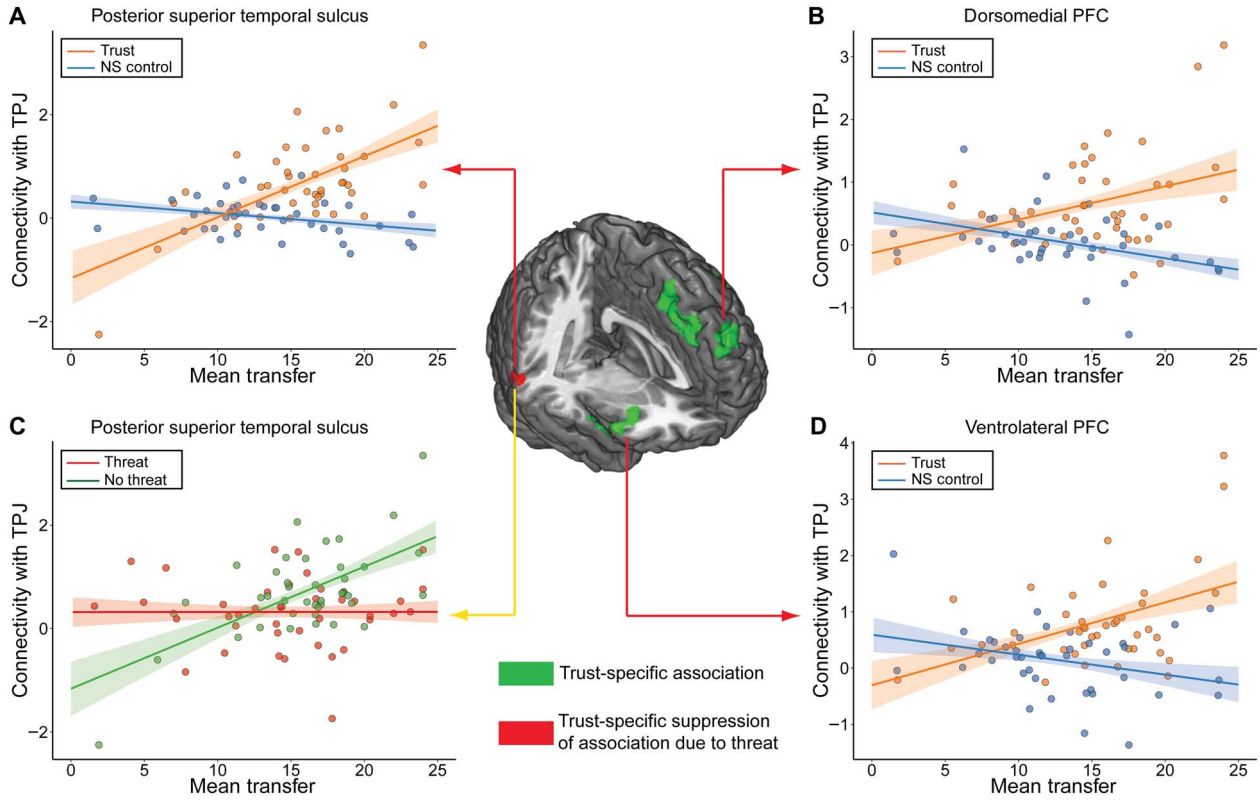

Fig. 3. TPJ functional connectivity is associated with transfer rates in the trust game. Trust-specific functional connectivity (**A**, **B**, and **D**) and threat-induced breakdown of connectivity (**C**) between the TPJ and a network of target regions. (A, B, and D) Trust-specific associations between transfer rate and trust-related neural activity reflecting the main effect of trust are shown in green activation clusters: Connectivity between the left TPJ and its targets is positively associated with trust (orange regression lines) but not with NS control decisions (blue regression lines) in the (A) bilateral posterior STS (pSTS; left peak at *xyz* = −62, −52, −5; right peak at *xyz* = 64, −43, 4), (B) dmPFC (peak at *xyz* = −12, 54, 40), anterior insula (left peak at *xyz* = −51, 21, −6; right peak at *xyz* = 56, 18, 1), and amygdala (peak at *xyz* = 28, 2 , −20). In contrast, mean transfers (i.e., investments) during the 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 NS control task (see table S4A for ROI analyses). (C) The results from the interaction contrast reflecting a trust-specific breakdown of the association between mean trust and TPJ connectivity are shown in the red activation cluster in STS (peak at *xyz* = 64, −43, 6; table S4B): Aversive affect causes a breakdown of the association between TPJ-pSTS connectivity and mean trust. 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, 6; see table S4C). Specifically, there is a positive association between TPJ-pSTS connectivity and the mean trust level when distortionary aversive affect is absent (green regression line), which is eliminated by threat (red regression line). This suggests that connectivity between the TPJ and its target region in pSTS supports general trust taking only in the absence of threat. The regression lines in (A to D) predict functional connectivity strength as a function of mean transfer levels based on an extended ordinary least squares model that estimates both the slope of the relationship between mean transfers and functional connectivity in the NS control task and the increase in this relationship in the trust task (relative to the NS control task). For this purpose, we extracted data from 6-mm spheres around individual interaction peak voxels (see Materials and Methods). Confidence bounds around regression lines reflect 95% confidence intervals around the model fit.

These results suggest that connectivity between the left TPJ and its target region in the contralateral pSTS supports trust when distortionary aversive affect is absent. However, in the presence of aversive affect, the relationship between the connectivity pattern in the trust network and behavioral trust was suppressed. Thus, aversive affect not only reduced average trust but also diminished specific relationships between the connectivity patterns in the TPJ network and behavioral indices of trust. Our results therefore suggest that the pSTS is a crucial neural node that mediates the breakdown of trust in the presence of threat.

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

The previous analyses indicate that aversive affect 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 affective reactions to the threatening context also had an impact on NS control decisions. We therefore addressed the question to what extent aversive affect impacts general choice-related neural circuitry in both the trust and the NS control game via an exploratory whole-brain analysis, investigating the main effect of aversive affect: ($trust_{threat}$ + NS $control_{threat}$) > ($trust_{no\ threat}$ + NS $control_{no\ threat}$). This identified a domain-general network of regions showing either suppression or enhancement in choice-related neural activity during both the trust and the NS control task (Fig. 4 and table S6). The suppression of neural activity in the threat condition (red time course; Fig. 4) relative to the no-threat condition (green time course; Fig. 4) was observed in both the trust and NS control task (fig. S3) in the bilateral posterior dlPFC (left: −62, −4, 18; k = 686; right: 62, −6, 28; k = 515), left amygdala (−24, −15, −23;

k = 226), posterior paracentral lobule (4, −36, 69; k = 309), left vlPFC (−48, 41, −8; k = 281), and vmPFC (−10, 44, −8; k = 464; table S6A). Significant enhancement of activity during decision-making under aversive affect (fig. S4) was obtained in the cerebellum (−4, −46, −24; k = 549; table S6B). Together, these results identify a network of domain-general regions whose decision-related activity is significantly affected by incidental aversive affect. 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 affect occur above and beyond domain-general effects on decision-making.

DISCUSSION

Incidental aversive affect is a ubiquitous phenomenon that pervades many aspects of human behavior and human social interaction. Here, we investigated the behavioral and neural impact of incidental affect on trust decisions. We used an experimental technique that establishes aversive affect 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 NS decision-making and was associated with consistent self-reports of the experience of aversive affect. We observed that aversive affect significantly reduced participants' trust in their partners. To the extent to which aversive affect is associated with stress, this result is consistent with a recent behavioral study that showed that acute stress reduces trust (*24*). Despite the fact that aversive affect was incidental to the decisions made by participants, we observed significant behavioral and neural effects of the aversive affective state.

Our behavioral results underline the importance of affect in decision-making even when they are unrelated to decision outcomes. As outlined previously (*8*, *36*), these findings contradict the assumption of consequentialist economic models that emotions can at most affect choices by changing participants' preferences over outcomes (*8*). Given that incidental emotions are by definition unrelated to the consequences of ongoing decisions, they are assumed to exert no influence on decision-making in these models. Further theoretical work may be needed to account for the impact of incidental affect on economic preferences, such as already proposed for anticipatory anxiety (*37*).

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 affect. While the TPJ was preferentially engaged during trust decisions, aversive affect led to a trust-specific breakdown of this

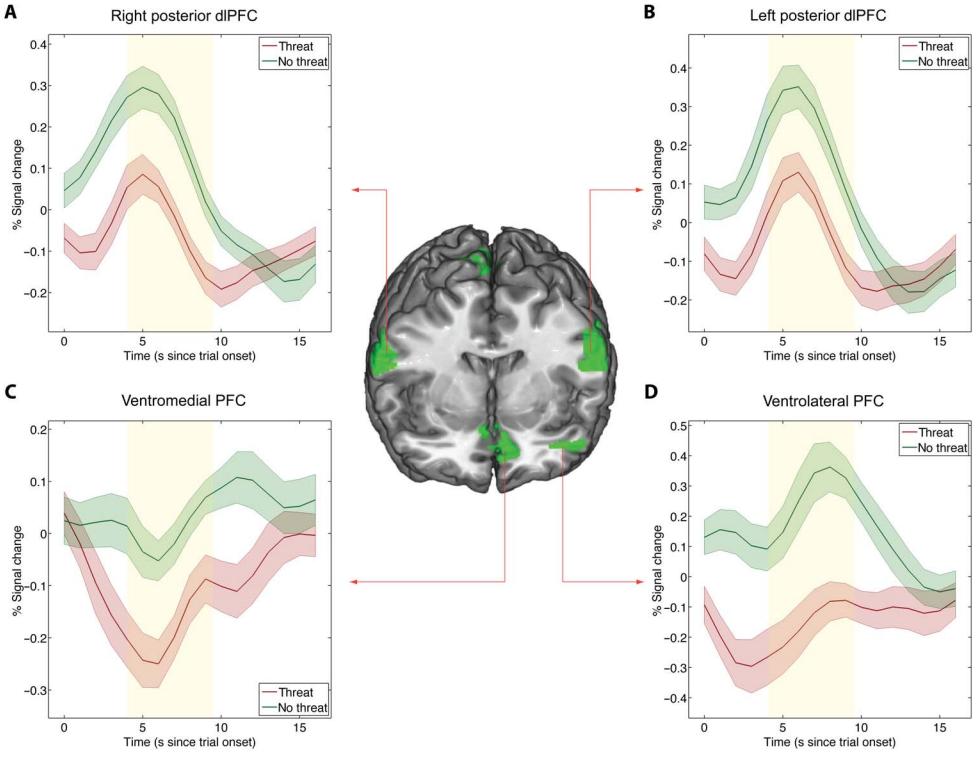

Fig. 4. The impact of aversive affect on choice domain–independent neural correlates of decision-making. We tested the main effect of aversive affect on the neural correlates of decision-making independent of the choice domain (social and NS). This analysis revealed a domain-general network consisting of the bilateral posterior dlPFC [right peak at *xyz* = −62, −4, 18 (**A**); left peak at *xyz* = 62, −6, 28 (**B**)] and two clusters in the vmPFC [left peak at *xyz* = −10, 44, −8; right peak at *xyz* = 6, 21, −14 (**C**)] and left vlPFC [peak at *xyz* = −48, 41, −8 (**D**)]. These regions show significant threat-related suppression (no threat > threat; regions shown in green) in choice-related activity during both trust and NS control trials. Additional regions that are shown in fig. S3 include the left amygdala (B) (peak at *xyz* = −24, −15, −23) and posterior paracentral lobule (peak at *xyz* = 4, −36, 69). Time courses reflect choice domain–independent activity that shows suppressions due to the aversive affect during decisions in both trust and NS control trials. To illustrate the equivalent effect of aversive affect, figs. S3 and S4 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-s choice period is displayed in yellow and is shifted for 4 s to account for the hemodynamic lag.

activation pattern (Fig. 2, A and B) in the left TPJ (and in right TPJ at a reduced threshold). Moreover, in the absence of aversive affect, we observed significant connectivity between the TPJ and amygdala during trust decisions, but this connectivity was disrupted when we induced aversive affect (Fig. 2, C and D). Aversive affect also disrupted the relation between the connectivity patterns in the neural network underlying trust and the magnitude of behavioral trust. In particular, functional connectivity strength with the TPJ was associated with mean trust levels for a network of regions consisting of the amygdala, dmPFC, STS, and vlPFC (tables S3 and S5 and Fig. 3, A, B, and D). Aversive affect caused a specific breakdown of this association for the pSTS, such that the connectivity between left TPJ and right pSTS no longer predicted overall trust-taking (Fig. 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 affected by incidental aversive affect.

The previous literature (*2*, *3*) has identified betrayal aversion as one of the key determinants of trust decisions in the trust game. Betrayal aversion means that participants 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 affective significance of these responses. These processes critically involve participants' mentalizing faculties and the assessment of the (negative) affective value of the possibility of being cheated by an untrustworthy partner (*16*–*18*).

The TPJ and the dmPFC have repeatedly been implicated in representing and interpreting others' mental states and behavior (*19*), and the amygdala has been shown to respond strongly to variations in the trustworthiness of faces (*30*) and other emotionally salient stimuli (*38*). We therefore conjectured that these regions also play a key role in the computations involved in assessing and evaluating the partner's anticipated responses in the trust game. This hypothesis is consistent with previous reports of TPJ, dmPFC, and amydgala activation during trust decisions (*5*, *16*–*18*). Our present results significantly extend these previous findings by 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 the TPJ and regions important for social cognition (dmPFC and STS) and emotion processing (amygdala and vlPFC) relate to each individuals' levels of trust.

However, we show that key components of these trust-supportive and trust-specific neural networks are suppressed by aversive affect. In particular, we find that threat of shock leads to (i) specific reductions of TPJ activation during trust decisions, (ii) specific reductions in the connectivity between the TPJ and amygdala during trust decisions, (iii) general reductions in choice-related activity in the amygdala, and (iv) trust-specific disruptions of the association between TPJ connectivity and mean trust. These results thus suggest that aversive affect undermines 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 in the change in TPJ-amygdala connectivity due to threat of shock: In the absence of threat, the 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 (fig. S3), and this preoccupation with the immediate and emotionally highly salient threatening context seems to prevent it from communicating with the TPJ. That is, aversive affect seems to reduce a participant's capacity to mentalize about and evaluate the emotional consequences of various trust levels, and as a consequence, participants reduce their behavioral trust toward their partner.

At the behavioral level, we observed that aversive affect reduced both trust and NS control decisions. This result is not expected, given that the fMRI design was conceptualized and optimized to reveal the effects of aversive affect on the neural correlates of trust decisions, which required the NS control task to be equivalent to the trust game in all respects, except that a randomization procedure and not another human being decided the back-transfer amount. This finding therefore underlines that the induction of aversive affect via threat of shock was successful. The comparable impact of aversive affect on trust and NS control trials furthermore confirms that our NS condition constitutes a well-matched control for the trust game.

Despite the similar behavioral effect of aversive affect on social and control decision, we identify a differential impact of aversive affect on the neural correlates of trust and NS decision-making. Our neuroimaging results therefore demonstrate a dissociation between the mental processes underlying trust and control decisions above and beyond the behavioral effects. This underlines the added explanatory value of measuring brain activity concurrent with behavior, because the neural results can reveal the presence of distinct cognitive and affective processes across experimental conditions that could remain concealed in experiments reporting only behavioral results. Specifically, we show an interaction between game type and aversive affect in TPJ activity and connectivity (Fig. 2), demonstrating that aversive affect impacted on social cognitive processes specifically during trust decisions but not during control decisions. Moreover, our neural results show that changes in the connectivity between social cognition regions (TPJ-dmPFC; TPJ-pSTS) are associated with trust decisions but not with decisions that do not involve an interaction partner (Fig. 3). Our neuroimaging findings therefore suggest that different mental processes underlie decisions across the trust and NS control games and that aversive affect therefore impacted separable underlying cognitive processes. Further support for this view is given by additional behavioral analyses demonstrating subtle performance differences across game types. Specifically, we find (i) significant differences in average transfer rates and reaction times, (ii) similarly strong correlations within a game type, but significantly different correlations across game types (but within shock conditions), and (iii) an absence of a significant correlation between the threat-induced differences in transfer rates in the trust and control games (reported in section S2). These results are consistent with previous observations that two experimental manipulations can generate behavioral effects of similar size but still reflect influences on very different mental operations (e.g., trust versus general risk-taking) as evident by different brain activity (*39*–*41*). Moreover, if the purpose of the study is to understand these differences in terms of brain activity and connectivity, then it is desirable that there are no differences in behavioral responses across conditions [e.g., (*39*, *41*)].

A notable finding is the threat-related suppression of activity in emotion-related regions, such as the anterior insula and amygdala (Fig. 4 and fig. S3). While previous studies have typically found increased activity in these regions during anticipation of negative outcomes, such as losses [e.g., (*13*, *42*)] and electrical shocks [e.g.,

(*43*, *44*)], our manipulation of maintaining a prolonged period of threat (average block length, about 39 s) suppressed activity in a network of regions that includes the vmPFC, anterior insula, and amygdala (see table S6). This deviation from previous studies is likely due to the sustained nature of the emotional context in the current study, which differs markedly from the approach taken in previous studies [e.g., (*43*, *44*)]. These studies have used the application of short and discrete cues that are predictably paired with an aversive event, thereby inducing phasic effects on the neural correlates of fear (*45*). Our results agree with the argument that anxiety, induced via prolonged periods of threat of unpredictable aversive events, is distinguishable from fear, which occurs in anticipation of predictable and aversive events (*45*). One possibility to consider in future research is that fear and anxiety may be distinguishable at the neural level, such that affect-relevant brain areas may show activity increases in response to momentary negative emotions, such as fear, but suppressed activity in response to sustained negative affect, such as anxiety.

An alternative explanation for our observation of suppression in behavioral trust and its neural correlates in TPJ is that participants may have simply been distracted by the presence of threat. However, our finding of significantly faster RTs in the presence of threat makes this alternative interpretation unlikely, as this pattern of results is opposite to what would be expected if participants were distracted or under heavier cognitive load. Numerous studies have consistently shown that a hallmark of cognitive load (or distraction) is a slowing of RTs and that longer RTs are often used as a validation that an experimental load or distraction manipulation worked (*46*). Our findings that RTs were faster under threat are therefore much more consistent with the interpretation that participants were emotionally aroused under threat (*47*) and that this affective arousal affected social decisions and their neural correlates. In addition, participants' decisions may become more impulsive or inconsistent in the presence of threat. However, analyses of decision variability (fig. S1D) demonstrated no differences in the SD of transfer rates in the presence compared to the absence of threat, therefore contradicting this possibility.

While we were careful in using a well-matched control condition for the trust game by offering ambiguous lotteries that provided exactly the same investment options and probabilities of receiving a beneficial back transfer, a number of potential limitations are worth mentioning. These include potential differences in the complexity and expected returns across the two games. Specifically, trust decisions might be of higher complexity because no information was provided about the repayment probability in the trust game, while a probability range was provided in the control game. This was done to preserve the social nature of trust choices, as outlined in detail in the "Task description" section in Materials and Methods. However, two pieces of evidence lend support to the notion that differences in task complexity did not strongly influence the current results: (i) Additional behavioral analyses using decision latency as a proxy for choice complexity (reported in section S3) show that our behavioral and neuroimaging results are very unlikely to have been influenced by differences in choice complexity; (ii) if our neuroimaging findings were driven by this condition difference in explicit knowledge about back-transfer probability, then we would expect to see differential activations in regions that process outcome probability, such as the cingulate gyrus, anterior insula, striatum, and parietal parietal cortex and dlPFC [see neurosynth meta-analysis for "probability"; for review, see (*48*)]. Our neuroimaging results are largely inconsistent with this and instead support the view that social cognition regions, such as TPJ and dmPFC support mentalizing during trust decisions. The games might also differ in their subjective expected payouts (note that objective payouts were matched across the games), because our behavioral data show that participants invested more in the trust compared to the control game (Fig. 1D). If differences in the subjective expected payouts drove decisions differentially in the trust compared to the control game, then we would have expected activations in a network of regions that consistently track expected payouts for the main effect of game type, including the ventral striatum [see the meta-analysis in (*42*) and the vmPFC in (*49*); for meta-analysis, see (*33*)]. Our imaging results are again inconsistent with this interpretation, showing activations in the TPJ and brain-behavior relationships in the dmPFC and pSTS. Note, however, that our neuroimaging results provide only post hoc and indirect evidence for the notion that differences in task complexity and subjective expected returns contributed little to the observed results. While the magnitude of any potential effects of these factors on trust decisions could be systematically tested by future research, our results support the view that trust decisions reflect social cognitive processes that are suppressed in the presence of threat. An interesting question arising from our findings is to what degree these effects of aversive affect on socio-cognitive processes may generalize to other domains of social and economic behavior [for instance, see (*50*, *51*)].

In conclusion, we report results that show a significant behavioral impact of incidental affect on trust-taking, and we identify the trust-specific neural mechanisms associated with the impact of aversive affect on trust. These effects were observed although the induced affect was unrelated to the choice outcomes in our task, confirming that incidental affect can have a powerful impact on behavior and its underlying mental operations. Our findings inform the development of economic and social theory and call for the integration of incidental affect in behavioral models of social and NS decision-making (*8*–*10*), such as done recently for anticipatory anxiety (*37*). In addition, by identifying the specific distortions of the neural network activity supporting trust, we provide a first step toward neural models that help us better understand these distortions. In particular, our results support the notion that an important mechanism through which aversive incidental affect 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 TPJ, dmPFC, and STS) 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 on the one hand and distortions of social behavior including trust on the other (*17*, *52*), our results may also be useful for understanding the neural circuitry associated with affect-related distortions of social behavior in psychiatric diseases.

MATERIALS AND METHODS

Participants

Forty-one human volunteers [mean age (SD), 22 (2.145); 17 females] from various Universities in Zürich participated in the current experiment. Only right-handed participants between the ages of 18 and 45 with no previous 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 simulations (*53*). All participants gave written informed consent to procedures

approved by the local ethics committee (Kantonale Ethikkommission, Zürich, Switzerland) before participating in the study. Participants 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 participants understood all aspects of the experiment. To this end, participants 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 participant's answers was confirmed by the experimenters and discussed in a brief interview that lasted for ca. 10 min. Participants 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 s per trial, understood the experimental setup, and were subsequently given the opportunity to ask further questions.

After completion of practice, participants were taken out of the scanner and washed their hands before placement of SCR and stimulation electrodes. Participants were then placed inside the scanner, and two ring electrodes were attached to the dorsum of the left hand: (i) The electrode providing relatively higher intensity stimulation was placed between 1 and 2 cm below the second carpometacarpal joint, and (ii) the electrode providing relatively lower intensity stimulation was placed 1 to 2 cm below the fifth carpometacarpal joint. To determine individual thresholds for high- and low-intensity stimulation, we followed a standard procedure (*50*) and used a visual analog rating scale with end points 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, 5 V, 50 mA; 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 the 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, was controlled by Cogent 2000 (www.vislab.ucl.ac.uk/cogent.php).

Task description

To investigate the effect of incidental affect on trust-taking, we used a hybrid fMRI design, in which aversive affect was manipulated in a blocked fashion while social (trust) and NS (control) tasks were presented in an event-related fashion. Specifically, we varied aversive affect by creating an expectancy of weak or strong unpredictable electrical stimulation (the duration for both was 20 ms) 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 (Fig. 1A). Stimulation intensity was communicated to participants in three ways: (i) via a verbal cue embedded in the 750-ms block cue ["strong" for treatment ("threat" condition) and "weak" for control ("no-threat" condition)]; (ii) via a predictable tactile reminder cue presented 700 ms after visual cue onset for a duration of 20 ms, which reflected the exact stimulation intensity of the current block; and (iii) via a specific background color that was consistently associated with either threat or no-threat blocks for each participant (color was counterbalanced across participants) 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 participants to augment the efficacy of the threat-of-shock treatment. For this purpose, the number of stimulation events was determined for each block by random draw from a γ distribution (shape parameter, 1; scale parameter, 1). Participants therefore experienced exactly one predictable reminder and, on average, one additional unpredictable electrical stimulation per block. 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 participant to maximize identification of the effects of aversive affect 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 (NS control or trust) and the level of stimulation (weak and strong) to be expected by participants for the rest of the block. After a brief and jittered interstimulus interval of 3 to 9 s, the first of three trials within a given block was displayed. In both the trust and the control game, participants were presented with a multiple-choice scenario, in which one of five amounts between 0 and 24 CHF could be transferred to player B or invested in an ambiguous lottery. In both games, participants faced an ambiguous back-transfer likelihood: In the trust game, participants were not informed about the probability of a beneficial back transfer and needed to infer this themselves for each interaction with a trustor. This was performed to avoid biasing participants' decisions and to ensure that participants did not simply focus on explicit repayment probabilities, thereby maintaining the social nature of interactions in the trust game. Ambiguity was also present in the NS control game, which provided only a probability range within which a beneficial back transfer could occur. To encourage investments into the lottery, the probability range used in the control game fell between 40 and 60%. This range was based on previous research that systematically varied the amount of ambiguity, demonstrating significant decreases in investments with increasing levels of ambiguity (*54*). It was matched to the probability of receiving a beneficial back transfer in the trust game, as it included the likelihood of receiving a back-transfer amount equal to or larger than the investment in the trust game (0.46 in our prerecorded trust game data), as well as participants' expectations about encountering a trustworthy trustee as identified by previous research (*2*). Participants always had the option to either invest all (24 CHF) or none (0 CHF) of their endowment. Moreover, each trial presented a novel choice scenario by (i) varying the intermediate options among the low (4, 6, or 8 CHF), medium (10, 12, or 14 CHF), and high (16, 18, or 20 CHF) categories of intermediate transfer amounts; (ii) varying the location of each choice option; and (iii) varying the location of the originally highlighted choice option. This variability was introduced to ensure that participants 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. Participants 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 four-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 to 9 s), a new trial began. Please note that, to control for wealth effects, participants in our experiment did not receive trial-by-trial feedback about the financial outcome of their choices in both the trust and the NS control game. Using one-shot games with no feedback, we precluded learning- and outcome- related signals commonly observed in valuation regions [e.g., (*55*, *56*)]. Participants completed 28 blocks (seven blocks per condition with an average length of 38.75 s) with three trials each in two runs. Thereafter, generic feedback was provided about the number of times player B and the lottery returned more or less than the participant's investment, and participants completed an additional 28 blocks (data not reported here).

Payment determination

We collected trustee responses in separate behavioral sessions that were conducted before 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 participants (investors) thus played against the prerecorded strategies of the trustees, i.e., a participant'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 participants 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. Last, to maintain the interactive nature of the trust game, we informed our participants 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. Results from an exit questionnaire assessing participants' perception of the social nature of their interactions during the main experiment indicate that participants trusted our experimental instructions and believed that they were interacting with a real partner who was determining their back-transfer amounts (as reported in fig. S1E, all ratings were significantly greater than the midpoint of the rating scale, $P < 0.0001$). During the experiment, the participants 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 participants selected two trials at random by dice throw, and payment was determined according to the decisions made by the participant and the trustee on the selected trials. 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, then the investor's payout was determined on the basis of the amount transferred to the trustee and the back-transfer amount of the specific trustee the investor was paired with on that trial (payout investor = 24 − transfer to trustee + back transfer from trustee; payout trustee = 24 + transfer from investor × 3 − back transfer to investor). If a control game was randomly chosen, then the computer algorithm randomly drew a payout amount from the distribution of trustees' back-transfer amounts. Our procedure therefore created equivalent payout amounts and likelihoods for the trust and control game.

Exit questionnaire

After completion of the experiment, participants filled out an exit questionnaire that probed their beliefs about the accuracy of our instructions, as well as affective reactions to our experimental manipulations. The main goal of the exit questionnaire was to measure whether participants believed our instructions. Note that we implemented these measurements routinely although we had little reason to believe that participants 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 participants. Participants were asked to rate seven 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 participants on the relevant trial. Participants' responses were entered into one-sample *t* tests testing whether responses were significantly greater than the midpoint of the scale (2, indicating neither believable nor unbelievable). Mean ratings for all statements were significantly greater than two, indicating that participants believed the statements [see fig. S1E; 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 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 adhesive straps after application of conductance gel. Participants' hands had been washed 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).

Preprocessing and statistical analyses of SCR data were performed using PsPM (PsychoPhysiological Modelling) (*57*). Because of technical problems, data from six participants included only one run (of two), and data from one participant were lost. Each participant's SCR data were downsampled to 10 Hz, low-pass filtered (cutoff frequency, 5 Hz), and subsequently *z*-transformed. Statistical analysis of the skin conductance data followed the approach commonly used in analyses of fMRI data. Specifically, multiple linear regression was used to estimate SCRs during decisions made in each of the task conditions, that is, during trust and NS control tasks and in the context of threat and no-threat treatment blocks. All events were modeled as Dirac delta functions and convolved using the canonical SCR function together with its temporal derivative (*58*). We took two precautions to ensure that electrical shocks did not influence estimates of arousal during the decision period: First, we modeled all shock events, entering them as four regressors that reflect the time points of strong or weak shock administration, separately for random (unpredictable) shocks and predictable shocks associated with the block cue. Second, we removed

from the regressors of interest all trials during which a shock occurred within an interval from 5 s before the onset of the decision screen until the button press. All these trials were added as two regressors of no interest, reflecting decisions made in the presence of strong and weak shocks. The statistical model therefore included a total of 11 regressors that reflected the onset times of decision screens in trust and NS control trials under expectancy of strong and weak electrical shocks (four regressors of interest during which no shock occurred), cue times indicating the onset of a block, delivery times of tactile stimulation (with four separate regressors including all predictable strong and weak shocks and all unpredictable strong and weak shocks), and the regressors of no interest for decision trials during which a weak or a strong shock occurred. Regressor estimates (β weights) for each condition were then used in follow-up analyses (one-sample *t* tests contrasting the difference between threatening and nonthreatening conditions and strong versus weak shocks) reported in Results.

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 eight-channel Philips sensitivity-encoded (SENSE) head coil. Structural 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 (EPI) sequence with an acceleration factor of 2.0. We acquired 45 axial slices covering the whole brain with a slice thickness of 2.8 mm (interslice gap, 0.8 mm; sequential acquisition; repetition time, 2470 ms; echo time, 30 ms; flip angle, 82°; field of view, 192 mm; matrix size, 68 by 68). To optimize functional sensitivity in the orbitofrontal cortex and medial temporal lobes, we used a tilted acquisition in an oblique orientation at 15° relative to the anterior commissure–posterior commissure 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. Last, functional data underwent spatial smoothing using an isotropic 6-mm full-width at half maximum Gaussian kernel.

Statistical analyses were carried out using the general linear model (GLM). Regressors of interest were modeled using a canonical hemodynamic response function (HRF) with time and dispersion derivatives to account for subject-to-subject and voxel-to-voxel variation in response peak and dispersion. Since our main interest was the impact of aversive affect on trust-taking, we modeled the decision period for the full RT on each trial, that is, from the onset of the decision screen until participants pressed the confirm button. This approach implicitly controls for differences in decision latencies (*59*). This was performed in the following four conditions: (i) trust game during relatively high-intensity stimulation expectancy (threat condition), (ii) trust game during relatively low-intensity stimulation expectancy (no-threat condition), (iii) control game during relatively high-intensity stimulation expectancy (threat condition), and (iv) control game during relatively low-intensity stimulation expectancy (no-threat condition). Last, the following regressors of no interest were included in our model: the actually realized weak and strong tactile stimulations during a block (modeled as two separate regressors, one including all the predictable and unpredictable weak and one including all the predictable and unpredictable strong shocks), 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 affect on the neural correlates of trust decisions. Trust-specific neural effects of aversive affect 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 NS 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 and control), threat (absent and present), and participant. We were particularly interested in trust-specific affect-induced suppression of activity and connectivity, which we tested via the interaction contrast ($trust_{no\ threat} > trust_{threat}$) > ($NS\ control_{no\ threat} > NS\ control_{threat}$) in the context of the flexible factorial design. A covariate reflective of each participant'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.

Our main analyses rely on a priori ROIs as we expected regions commonly implicated in the major cognitive and affective component processes of trust to be affected by aversive affect. Specifically, we hypothesized that participants 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 (*20*). To identify regions implicated in theory of mind, we consulted neurosynth.org (*31*), which offers a means to obtain automated meta-analyses over a large number of previous 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 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 the 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 (*19*). Furthermore, because of its prominent role in signaling emotional salience (*11*), 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 (automated anatomical labeling) atlas implemented in WFU Pickatlas. The following independent ROI masks were created via automated meta-analyses from neurosynth.org: (i) bilateral TPJ (neurosynth term: theory of mind), with peak voxels in the left (−60,−56,14) and right TPJ (56,−58,20) and sizes

of 1031 and 1416 voxels, respectively. Note that the ventral part of this mask also contains voxels from pSTS. For simplicity, we refered to this mask nevertheless as "TPJ". (ii) dmPFC (neurosynth term: theory of mind), with a peak voxel in the 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 the AAL atlas using an anatomical mask for bilateral amygdala 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 the ventral striatum and vmPFC (*60*) (neurosynth term: reward), as well as a region implicated recently in one-shot trust decisions by a recent meta-analysis (*28*) 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; vmPFC (neurosynth search term: ventromedial) with a peak voxel in medial vmPFC (8, 24, −12), with a size of 1327 voxels; left anterior insula with peak voxels at −42, 18, 2 and a size of 13,844 voxels; and right anterior insula with peak voxels at 42, 18, 2 and a size of 13,871 voxels.

Furthermore, to identify whether extended networks outside our ROIs show effects of interest, we conducted exploratory whole-brain analyses at an FWE-corrected extent threshold of $P < 0.05$ ($k > 226$, initial cluster-forming height threshold $P < 0.001$). Last, to characterize activation patterns of interest, such as time courses and activation differences due to aversive affect, regression coefficients (beta weights) for the canonical HRF regressors were extracted with rfxplot (*61*) from 6-mm spheres around individual participants' peak voxel that showed significant effects of interest on BOLD (Blood-oxygen-level dependent) 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

PPI analyses were conducted using the generalized form of context-dependent PPIs toolbox (gPPI toolbox (*35*), using the same statistical model as outlined above. All voxels that survived SV FWE correction for the interaction contrast in the left TPJ (−60, −54, 19; $k = 95$) were used as a seed region (shown in blue color in Fig. 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. Psychological interaction regressors for each of the task type and stimulation intensity combinations [control decisions during (i) weak and (ii) strong stimulation and trust decisions during (iii) weak and (iv) strong stimulation] were created by multiplying the estimated neural activity during the relevant decisions with condition-specific onset and offset times convolved with the canonical HRF. A new GLM (general linear model mentioned above) was then estimated for each participant 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 affect 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 and NS control), threat (absent and 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 NS control task in the absence of threat ($\text{Trust}_{\text{no threat}} > \text{NS Control}_{\text{no threat}}$). 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 NS 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 illustrated these results in Fig. 3 by regression plots generated with coefficients reflecting functional connectivity strength for each of the conditions and 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)

$$y_{ik} = \beta_0 + \beta_1 \text{Transfer}_{ik} + \beta_2 \text{GameType}_k + \beta_3 \text{Threat}_k + \beta_4(\text{Transfer}_{ik} \times \text{GameType}_k) + \beta_5(\text{Transfer}_{ik} \times \text{Threat}_k) + \beta_6(\text{GameType}_k \times \text{Threat}_k) + \beta_7(\text{Transfer}_{ik} \times \text{GameType}_k \times \text{Threat}_k) + \varepsilon_{ik}$$

The dependent variable y_{ik} is the functional connectivity strength between a given brain region and the TPJ for individual i in GameType_k. Transfer_{ik} is the mean amount sent by individual i in GameType_k. GameType is a dummy variable encoding whether decisions were made in the trust or the control task (1 indicates trust, and 0 indicates NS control task). Threat is a dummy variable encoding whether decisions were made in the presence or absence of threat (1 indicates presence, and 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 Fig. 3, A, B, and D), and the sum of the coefficients for Transfer_{ik} (β_1) and the interaction term between $\text{Transfer}_{ik} \times \text{GameType}_k$ (β_4) measures the trust-related slope increase in the relationship between TPJ connectivity and mean transfers in the absence of threat (see orange lines in Fig. 3, A, B, and D). 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 $\text{Transfer}_{ik} \times \text{GameType}_k$ (β_4), the interaction term between $\text{Transfer}_{ik} \times \text{Threat}_k$ (β_5), and the interaction term between $\text{Transfer}_{ik} \times \text{GameType}_k \times \text{Threat}_k$ (β_7) measures the slope of the relationship between TPJ connectivity and mean trust in the presence of threat (see red line in Fig. 3C).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/3/eaau3413/DC1

Section S1. Ordinary least squares regression investigating the influence of experienced electrical stimulation on choice

Section S2. Behavioral differences between trust and control decisions

Section S3. Assessing complexity differences across game types using choice latency

Section S4. Assessing the lateralization of neuroimaging results

Fig. S1. Manipulation checks.

Fig. S2. Additional post hoc inspection of the significant interactions reported in the main paper within all voxels of independent TPJ and amygdala masks.

Fig. S3. Main effect of threat 1: Suppression of game type–independent neural correlates of decision-making.

Fig. S4. Main effect of threat 2: Enhancement of game type–independent neural correlates of decision-making.
Table S1. Ordinary least squares regression results reflecting the influence of experienced electrical stimulation on choice behavior.
Table S2. ROI analyses investigating trust-specific neural correlates in regions associated with social cognition and valuation.
Table S3. ROI analyses investigating TPJ-amygdala connectivity patterns.
Table S4. ROI analyses investigating brain-behavior relationships.
Table S5. Whole-brain analyses investigating brain-behavior relationships.
Table S6. Whole-brain analyses investigating the main effect of threat.
Reference (*62*)

REFERENCES AND NOTES

1. M. R. Delgado, R. H. Frank, E. A. Phelps, Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat. Neurosci.* **8**, 1611–1618 (2005).
2. I. Bohnet, R. Zeckhauser, Trust, risk and betrayal. *J. Econ. Behav. Organ.* **55**, 467–484 (2004).
3. J. A. Aimone, D. Houser, B. Weber, Neural signatures of betrayal aversion: An fMRI study of trust. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **281**, 20132127 (2014).
4. B. King-Casas, C. Sharp, L. Lomax-Bream, T. Lohrenz, P. Fonagy, P. Read Montague, The rupture and repair of cooperation in borderline personality disorder. *Science* **321**, 806–810 (2008).
5. P. M. Gromann, D. J. Heslenfeld, A.-K. Fett, D. W. Joyce, S. S. Shergill, L. Krabbendam, Trust versus paranoia: Abnormal response to social reward in psychotic illness. *Brain* **136**, 1968–1975 (2013).
6. A. Cohn, J. Engelmann, E. Fehr, M. A. Maréchal, Evidence for countercyclical risk aversion: An experiment with financial professionals. *Am. Econ. Rev.* **105**, 860–885 (2015).
7. K. M. Harlé, A. G. Sanfey, Incidental sadness biases social economic decisions in the ultimatum game. *Emotion* **7**, 876–881 (2007).
8. G. Loewenstein, Emotions in economic theory and economic behavior. *Am. Econ. Rev.* **90**, 426–432 (2000).
9. E. A. Phelps, K. M. Lempert, P. Sokol-Hessner, Emotion and decision making: Multiple modulatory neural circuits. *Annu. Rev. Neurosci.* **37**, 263–287 (2014).
10. J. S. Lerner, Y. Li, P. Valdesolo, K. S. Kassam, Emotion and decision making. *Annu. Rev. Psychol.* **66**, 799–823 (2015).
11. K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, L. F. Barrett, The brain basis of emotion: A meta-analytic review. *Behav. Brain Sci.* **35**, 121–143 (2012).
12. J. W. Kable, P. W. Glimcher, The neurobiology of decision: Consensus and controversy. *Neuron* **63**, 733–745 (2009).
13. C. M. Kuhnen, B. Knutson, The neural basis of financial risk taking. *Neuron* **47**, 763–770 (2005).
14. S. Schulreich, Y. G. Heussen, H. Gerhardt, P. N. C. Mohr, F. C. Binkofski, S. Koelsch, H. R. Heekeren, Music-evoked incidental happiness modulates probability weighting during risky lottery choices. *Front. Psychol.* **4**, 981 (2014).
15. F. Krueger, K. McCabe, J. Moll, N. Kriegeskorte, R. Zahn, M. Strenziok, A. Heinecke, J. Grafman, Neural correlates of trust. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20084–20089 (2007).
16. T. Baumgartner, M. Heinrichs, A. Vonlanthen, U. Fischbacher, E. Fehr, Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* **58**, 639–650 (2008).
17. C. S. Sripada, M. Angstadt, S. Banks, P. J. Nathan, I. Liberzon, K. Luan Phan, Functional neuroimaging of mentalizing during the trust game in social anxiety disorder. *Neuroreport* **20**, 984–989 (2009).
18. F. Krueger, J. Grafman, K. McCabe, Neural correlates of economic game playing. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3859–3874 (2008).
19. F. Van Overwalle, Social cognition and the brain: A meta-analysis. *Hum. Brain Mapp.* **30**, 829–858 (2009).
20. J. K. Rilling, A. G. Sanfey, The neuroscience of social decision-making. *Annu. Rev. Psychol.* **62**, 23–48 (2011).
21. A. Schmitz, C. Grillon, 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).
22. R. Westermann, K. Spies, G. Stahl, F. W. Hesse, Relative effectiveness and validity of mood induction procedures: A meta-analysis. *Eur. J. Soc. Psychol.* **26**, 557–580 (1996).
23. M. Martin, On the induction of mood. *Clin. Psychol. Rev.* **10**, 669–697 (1990).
24. O. FeldmanHall, C. M. Raio, J. T. Kubota, M. G. Seiler, E. A. Phelps, The effects of social context and acute stress on decision making under uncertainty. *Psychol. Sci.* **26**, 1918–1926 (2015).
25. A. J. Porcelli, M. R. Delgado, Acute stress modulates risk taking in financial decision making. *Psychol. Sci.* **20**, 278–283 (2009).
26. O. J. Robinson, K. Vytal, B. R. Cornwell, C. Grillon, The impact of anxiety upon cognition: Perspectives from human threat of shock studies. *Front. Hum. Neurosci.* **7**, 203 (2013).
27. B. Wicker, C. Keysers, J. Plailly, J. P. Royet, V. Gallese, G. Rizzolatti, Both of us disgusted in My insula: The common neural basis of seeing and feeling disgust. *Neuron* **40**, 655–664 (2003).
28. G. Bellucci, S. V. Chernyak, K. Goodyear, S. B. Eickhoff, F. Krueger, Neural signatures of trust in reciprocity: A coordinate-based meta-analysis. *Hum. Brain Mapp.* **38**, 1233–1248 (2016).
29. J. LeDoux, The amygdala. *Curr. Biol.* **17**, R868–R874 (2007).
30. J. S. Winston, B. A. Strange, J. O'Doherty, R. J. Dolan, Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nat. Neurosci.* **5**, 277–283 (2002).
31. T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, T. D. Wager, Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
32. C. C. Ruff, E. Fehr, The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurosci.* **15**, 549–562 (2014).
33. O. Bartra, J. T. McGuire, J. W. Kable, The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).
34. L. Pessoa, J. B. Engelmann, Embedding reward signals into perception and cognition. *Front. Neurosci.* **4**, 17 (2010).
35. D. G. McLaren, M. L. Ries, G. Xu, S. C. Johnson, A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches. *Neuroimage* **61**, 1277–1286 (2012).
36. K. Wälde, A. Moors, Current emotion research in economics. *Emot. Rev.* **9**, 271–278 (2017).
37. A. Caplin, J. Leahy, Psychological expected utility theory and anticipatory feelings. *Q. J. Econ.* **116**, 55–79 (2001).
38. L. Pessoa, Emotion and cognition and the amygdala: From "what is it?" to 'what's to be done?'. *Neuropsychologia* **48**, 3416–3429 (2010).
39. G. Hein, Y. Morishima, S. Leiberg, S. Sul, E. Fehr, The brain's functional network architecture reveals human motives. *Science* **351**, 1074–1078 (2016).
40. J. P. O'Doherty, T. W. Buchanan, B. Seymour, R. J. Dolan, Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron* **49**, 157–166 (2006).
41. D. Wilkinson, P. Halligan, The relevance of behavioural measures for functional-imaging studies of cognition. *Nat. Rev. Neurosci.* **5**, 67–73 (2004).
42. B. Knutson, S. M. Greer, Anticipatory affect: Neural correlates and consequences for choice. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3771–3786 (2008).
43. M. R. Delgado, R. L. Jou, J. E. LeDoux, E. A. Phelps, Avoiding negative outcomes: Tracking the mechanisms of avoidance learning in humans during fear conditioning. *Front. Behav. Neurosci.* **3**, 33 (2009).
44. E. A. Phelps, M. R. Delgado, K. I. Nearing, J. E. LeDoux, Extinction learning in humans. *Neuron* **43**, 897–905 (2004).
45. M. Davis, D. L. Walker, L. Miles, C. Grillon, Phasic vs sustained fear in rats and humans: Role of the extended amygdala in fear vs anxiety. *Neuropsychopharmacology* **35**, 105–135 (2010).
46. H. Pashler, Dual-task interference in simple tasks: Data and theory. *Psychol. Bull.* **116**, 220–244 (1994).
47. M. M. Bradley, M. Codispoti, B. N. Cuthbert, P. J. Lang, Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion* **1**, 276–298 (2001).
48. M. L. Platt, S. A. Huettel, Risky business: The neuroeconomics of decision making under uncertainty. *Nat. Neurosci.* **11**, 398–403 (2008).
49. J. Gläscher, A. N. Hampton, J. P. O'Doherty, Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cereb. Cortex* **19**, 483–495 (2009).
50. J. B. Engelmann, F. Meyer, E. Fehr, C. C. Ruff, Anticipatory anxiety disrupts neural valuation during risky choice. *J. Neurosci.* **35**, 3085–3099 (2015).
51. C. J. Charpentier, C. Hindocha, J. P. Roiser, O. J. Robinson, Anxiety promotes memory for mood-congruent faces but does not alter loss aversion. *Sci. Rep.* **6**, 24746 (2016).
52. V. B. Gradin, A. Pérez, J. A. MacFarlane, I. Cavin, G. Waiter, J. Engelmann, B. Dritschel, A. Pomi, K. Matthews, J. D. Steele, Abnormal brain responses to social fairness in depression: An fMRI study using the ultimatum game. *Psychol. Med.* **45**, 1241–1251 (2015).
53. T. Yarkoni, T. S. Braver, Cognitive neuroscience approaches to individual differences in working memory and executive control: Conceptual and methodological issues, in *The Handbook of Individual Differences in Cognition*, A. Gruszka, G. Matthews, B. Szymura, Eds. (Springer, 2010), pp. 87–107.

54. A. Tymula, L. A. R. Belmaker, A. K. Roy, L. Ruderman, K. Manson, P. W. Glimcher, I. Levy, Adolescents' risk-taking behavior is driven by tolerance to ambiguity. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17135–17140 (2012).
55. D. S. Fareri, L. J. Chang, M. R. Delgado, Computational substrates of social value in interpersonal collaboration. *J. Neurosci.* **35**, 8170–8180 (2015).
56. T. E. J. Behrens, L. T. Hunt, M. W. Woolrich, M. F. S. Rushworth, Associative learning of social value. *Nature* **456**, 245–249 (2008).
57. D. R. Bach, G. Flandin, K. J. Friston, R. J. Dolan, Time-series analysis for rapid event-related skin conductance responses. *J. Neurosci. Methods* **184**, 224–234 (2009).
58. D. R. Bach, K. J. Friston, R. J. Dolan, An improved algorithm for model-based analysis of evoked skin conductance responses. *Biol. Psychol.* **94**, 490–497 (2013).
59. J. Grinband, T. D. Wager, M. Lindquist, V. P. Ferrera, J. Hirsch, Detection of time-varying signals in event-related fMRI designs. *Neuroimage* **43**, 509–520 (2008).
60. D. J. Levy, P. W. Glimcher, The root of all value: A neural common currency for choice. *Curr. Opin. Neurobiol.* **22**, 1027–1038 (2012).
61. J. Gläscher, Visualization of group inference data in functional neuroimaging. *Neuroinformatics* **7**, 73–82 (2009).
62. B. R. Cornwell, A. M. Echiverri, M. F. Covington, C. Grillon, Modality-specific attention under imminent but not remote threat of shock: Evidence from differential prepulse inhibition of startle. *Psychol. Sci.* **19**, 615–622 (2008).

Acknowledgments: We thank K. Treiber and T. Williams for help with data collection and C. Rorden and A. Etter for coding support. **Funding:** J.B.E. acknowledges support from Amsterdam Brain and Cognition (ABC), the Radboud Excellence Initiative, the NCCR Affective Sciences, and the Mercator Foundation Switzerland. C.C.R. received funding from the Swiss National Science foundation (grant no. 100019L_173248) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant no. 725355, ERC consolidator grant BRAINCODES). E.F. acknowledges funding from the European Research Council [ERC grant no. 295642 on the Foundations of Economic Preferences (FEP)]. **Author contributions:** J.B.E., C.C.R., F.M., and E.F. designed the study. J.B.E. and F.M. collected the data. J.B.E. analyzed the data. C.C.R. and E.F. provided supervision. J.B.E., C.C.R., and E.F. wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Codes and behavioral and psychophysiological data presented in this article are available at figshare.com (10.6084/m9.figshare.6148295). The fMRI data presented in this article are available at neurovault.org (www.neurovault.org/collections/3736). Additional data related to this paper may be requested from the authors.

Submitted 31 May 2018
Accepted 29 January 2019
Published 13 March 2019
10.1126/sciadv.aau3413

Citation: J. B. Engelmann, F. Meyer, C. C. Ruff, E. Fehr, The neural circuitry of affect-induced distortions of trust. *Sci. Adv.* **5**, eaau3413 (2019).