Neuron
Article

The Neural Circuitry of a Broken Promise

Thomas Baumgartner,[1,6,*] Urs Fischbacher,[2,3,6] Anja Feierabend,[1] Kai Lutz,[4] and Ernst Fehr[1,5,*]
[1]Institute for Empirical Research in Economics, Laboratory for Social and Neural Systems Research, University of Zurich, Switzerland
[2]Department of Economics, University of Konstanz, Germany
[3]Thurgau Institute of Economics, Kreuzlingen, Switzerland
[4]Institute of Psychology, Department of Neuropsychology, University of Zurich, Switzerland
[5]Collegium Helveticum, Switzerland
[6]These authors contributed equally to this work
*Correspondence: t.baumgartner@iew.uzh.ch (T.B.), efehr@iew.uzh.ch (E.F.)
DOI 10.1016/j.neuron.2009.11.017

SUMMARY

Promises are one of the oldest human-specific psychological mechanisms fostering cooperation and trust. Here, we study the neural underpinnings of promise keeping and promise breaking. Subjects first make a promise decision (promise stage), then they anticipate whether the promise affects the interaction partner's decision (anticipation stage) and are subsequently free to keep or break the promise (decision stage). Findings revealed that the breaking of the promise is associated with increased activation in the DLPFC, ACC, and amygdala, suggesting that the dishonest act involves an emotional conflict due to the suppression of the honest response. Moreover, the breach of the promise can be predicted by a perfidious brain activity pattern (anterior insula, ACC, inferior frontal gyrus) during the promise and anticipation stage, indicating that brain measurements may reveal malevolent intentions before dishonest or deceitful acts are actually committed.

INTRODUCTION

The human capacity to establish and enforce social norms is one of the decisive reasons for the uniqueness of human cooperation in the animal kingdom (Fehr and Fischbacher, 2003). Such norms constitute standards of behavior that are based on widely shared beliefs on how individuals ought to behave in a given situation (Ellickson, 2001; Elster, 1989; Horne, 2001; Voss, 2001). In modern human societies, a large cooperative infrastructure in the form of laws, impartial courts, and the police exist, which ensure that cooperative agreements, for example in the form of enforceable contracts, are kept (Fehr et al., 2002). However, it is obvious that in more than 90 percent of human history no such cooperative infrastructure existed. Thus, in ancient times, other more basic forms of cooperative agreements must have evolved in order to foster trust, cooperation, and partnership formation. One basic form of such cooperative agreements are promises, which might in fact constitute the precursor of enforceable contracts in contemporary times. Promises constitute oral and "nonbinding" cooperative agreements, which have the goal to strengthen the belief in the exchange partner that one can be relied upon (Charness and Dufwenberg, 2006). Despite their nonbinding nature, many everyday social and economic exchange situations are still based on such oral promises. However, although important work examining the neural basis of social cooperation (Baumgartner et al., 2008a; Behrens et al., 2008; Delgado et al., 2005; King-Casas et al., 2005; Rilling et al., 2002, 2007; Singer et al., 2006; Tabibnia et al., 2008), social comparison, and competition (Decety et al., 2004; Delgado et al., 2008; Fliessbach et al., 2007; Zink et al., 2008), as well as social punishment and norm violations (Buckholtz et al., 2008; de Quervain et al., 2004; Eisenberger et al., 2003; Knoch et al., 2006, 2008; Meyer-Lindenberg et al., 2006; Sanfey et al., 2003; Spitzer et al., 2007) exists, the brain systems involved in nonbinding cooperative agreements still remain unknown. Studying the neural underpinnings of these nonbinding cooperative agreements is particularly interesting because promises not only can be kept, but also broken. In fact, material incentives to cheat are ubiquitous in human societies, and promises thus can also be misused in any kind of social or economic exchange situation between two or more individuals to cheat the exchange partner. Business people, politicians, diplomats, lawyers, and students in the experimental laboratory who make use of private information do not always do so honestly (Gneezy, 2005).

In real life, one reason for keeping promises is to facilitate the future cooperation of potential exchange partners. However, we also believe that humans often keep promises because this is "the right thing to do." Promises in this case are kept even in one-shot interaction, i.e., although the keeping of the promise implies a net cost to the promise keeper. In fact, decisive evidence from behavioral experiments reveals a preference for promise keeping in one-shot situations (Charness and Dufwenberg, 2006; Vanberg, 2008). Thus, it is possible to distinguish two major motivations behind promise keeping: first, *instrumental* promise keeping for the purpose of facilitating future cooperation, and second, *intrinsic* promise keeping for the purpose of "doing the right thing." In this paper, we focused on the second motivational source of promise keeping.

For that purpose, we applied a modified version of an economic trust game paradigm (Figure 1) where subjects were completely free to decide whether to keep or to break a promise and where keeping or breaking a promise caused real monetary consequences (either benefits or costs) for both exchange partners. In this economic trust game paradigm, two subjects

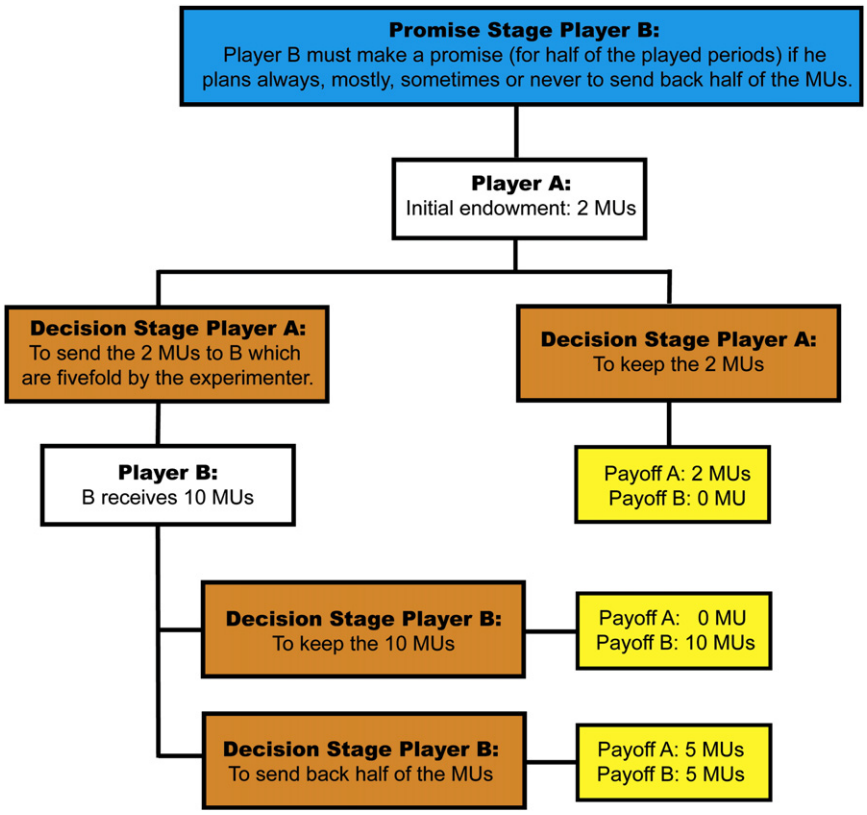

Figure 1. Trust Game with Antecedent Promise Stage

Depicted are the different stages of the economic trust game with antecedent promise stage. In the trust game used in the present study, two players A and B interact anonymously with each other during one trial. A receives an endowment of 2 money units (MUs) at the beginning of each trial, whereas B receives nothing. A has to make the first decision. He can send his endowment of two MUs to B (case 1), or he can keep his endowment (case 2). If A trusts B and sends his endowment (case 1), the experimenter increases the amount sent by a factor of five, so that B receives 10 MUs. At that moment, B has 10 MUs and A has nothing. B then has the choice of sending back nothing or half of the 10 MUs. Thus, if B acts trustworthily and sends back half, both players earn 5 MUs, but if B keeps all the money, he earns 10 MUs and A, who trusted B, earns nothing. In case 2, that is, if A does not trust B, A keeps his or her endowment of 2 MUs and B gets nothing. In total, 24 such trust game trials are played with different, randomly selected interaction partners. In half of the played rounds, B has to make a promise for three subsequent trials whether he *always*, *mostly*, *sometimes*, or *never* plans to send back half of the money. A is always informed about B's promise, and B can keep the promise, but he is also allowed to break it. Color coding: blue color, promise stage of player B; orange color, decision stages of either player A or B. Note that player A's decision stage is at the same time as player B's anticipation stage, during which player B is informed that player A is now deciding (see Figure 2); yellow color, outcome stage player A and B.

interacting anonymously are in the role of an investor (player A) and a trustee (player B). For the purpose of the study, we focused on the role of the trustee whose brain activity was measured in the brain scanner. The trustee first has to make a promise decision at the beginning of a series of three subsequent trust game trials, indicating whether he *always*, *mostly*, *sometimes*, or *never* plans to be trustworthy. In this context, being trustworthy means sharing the available money so that both players earn the same amount. Player A, the investor, is always informed about B's promise, and can then decide (based on B's promise) whether to trust him and invest money or whether not to trust him and thus to keep the initial endowment. In case player A trusts player B, which is almost only the case if player B chooses a high promise level (see Results), the experimenter increases the amount player A sends by the factor of five. Player B can then decide to keep the promise and thus honor an investor's trust by sending back half of the money, but he may also break the promise and thus violate the investor's trust by not sharing. The experiment consisted of four promise decisions with three subsequent trust game trials, meaning that subjects played a total of 12 trust game trials in the promise condition (i.e., with a promise stage). As a control condition, we also implemented 12 trust game trials *without* the opportunity of making a promise decision. The trustee thus faced a total of 24 trust game trials with 24 different, anonymous, and randomly selected interaction partners, half of the trials played with a promise stage and half of them without the opportunity to make a promise. Please note that the social interactions between trustees and interaction partners were genuine, that is, the trustees in the scanner faced the decisions of 24 real human interaction partners and their choices actually affected the interaction partners' monetary payoffs (please see Supplemental Experimental Procedures for details).

This design enables us to study three different processes that play an important role during nonbinding cooperative agreements: (1) the process of promising, (2) the process of anticipating the effect of the promise on the exchange partner's decision to trust, and (3) the decision-making process during which the decision to keep or to break the promise has to be implemented (see Figure 2 for two timelines of trust game trials with and without opportunity to make a promise). We are particularly interested in whether the brain activity pattern differs at the different stages of the paradigm dependent on the decision to keep or to break the promise.

In the experiment, the trustees were completely free to choose the strength of their promise (i.e., whether they promise *always*, *mostly*, *sometimes*, or *never* to share the money in the subsequent three trials) and to honor or break their promise. This led to two large behavioral clusters of individuals and only very few subjects did not belong to one of the two clusters. First, a

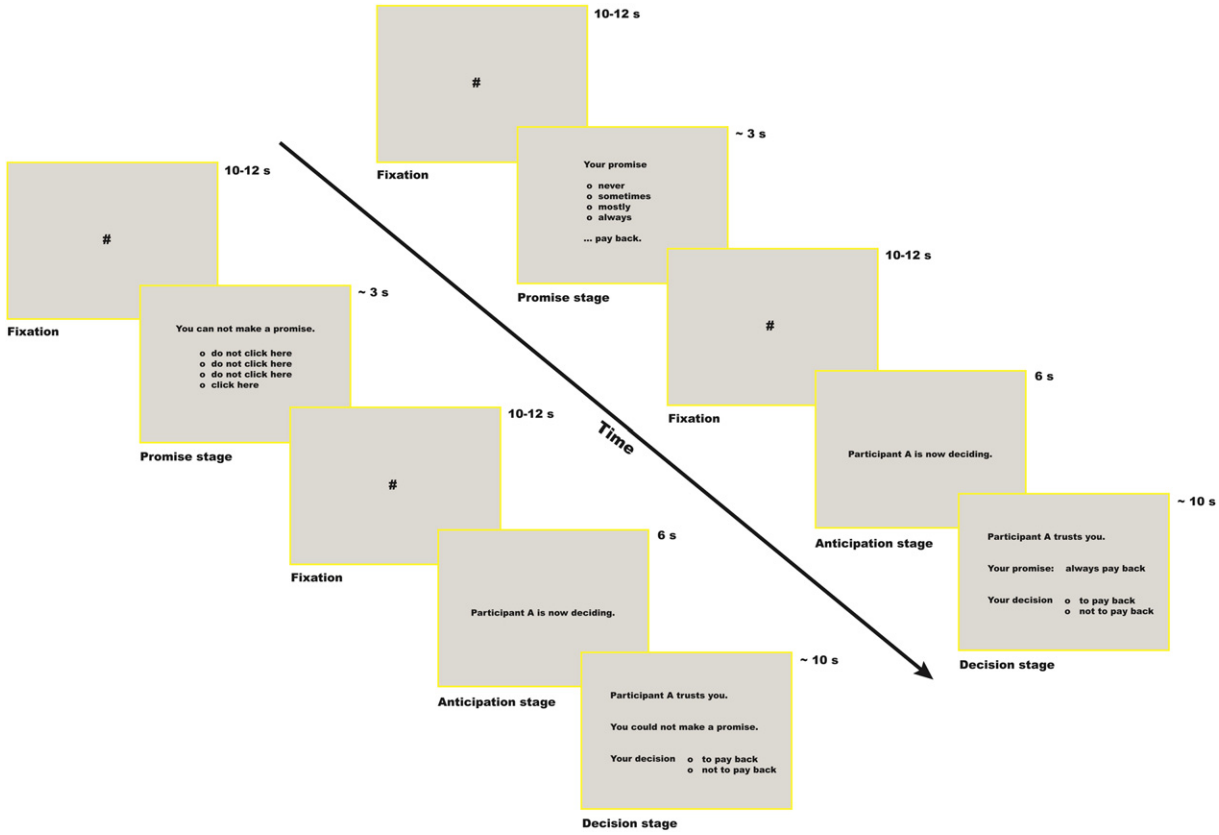

Figure 2. Timeline for Two Trials of the Trust Game with and without Antecedent Promise Stage

The trust game trials start with a fixation epoch that lasts for 10–12 s (randomly jittered). After this fixation epoch, the promise stage begins in 8 of 24 trust trials, during which the subject has to implement his promise level for three subsequent trust game trials (within a time restriction of 9 s, mean: ~3 s) or during which he receives the information that he cannot decide about a promise level. After the promise stage, there is another fixation epoch lasting for 10–12 s (randomly jittered). Then the anticipation stage begins, which last for 6 s, during which the subject is informed that his assigned player A is now deciding. This anticipation stage is followed by the decision stage, which is divided into three parts. First, the subject is informed for 6 s whether player A trusted him or not (not depicted). The subject is then reminded on the same decision screen of his promise or that he could not make a promise for the current trial. This information is presented for 3 s. Finally, after 9 s in total, the decision options are presented on the same screen, allowing the subject to implement his decision within a time restriction of 7 s. The first 6 s of the decision stage are referred to in the paper as decision phase A, whereas the second 3 s until button press are referred to as decision phase B (average response time from the beginning of decision phase A until button press: ~10 s). Finally, a trust game trial is completed by the profit stage (not depicted), which presents the outcome of both players for the current trust game trial for 6 s and provides the information that a new player A is assigned to the subject.

substantial proportion of the subjects promised to share the money "always" but actually did not share it in the subsequent trust games (dishonest subjects). Second, another large proportion of the subjects also promised to share the money "always" but these subjects subsequently kept their promise (honest subjects). These two clusters of individuals also behaved very consistently when they could not make a promise, with the dishonest subjects almost never sharing the money, while the honest subjects almost always shared the money (for detailed statistical information, please see the behavioral analyses in Results).

This behavioral data pattern requires that special care be taken in the analysis of the neuroimaging data in order to control for payoff differences and differences in fairness related behaviors. In particular, it is not possible to make simple, direct comparisons between the dishonest and the honest subjects' brain activity within the "promise possible" condition or within the "no promise possible" condition because such comparisons will be confounded with fairness differences and differences in material payoffs across the subjects. For this reason, we computed the following serial subtraction term for each of the stages of our paradigm: $[\text{Promise (P)} - \text{No Promise (NoP)}]^{\text{Dishonest subjects}} - [\text{Promise (P)} - \text{No Promise (NoP)}]^{\text{Honest subjects}}$, where (P) indicates the "promise possible condition" and (NoP) the "no promise possible" condition. Note that this contrast controls for fairness and payoff differences because dishonest subjects make the same unfair choices and earn the same payoff across the "promise possible" and the "no promise possible" condition. Thus, the brain activity in the contrast $(\text{P} - \text{NoP})^{\text{Dishonest subjects}}$ does not contain fairness and payoff-related brain activation. Likewise, honest subjects make the same fair choices and earn the same payoff across the "promise possible" and the "no promise possible" condition and, hence, the activity in the contrast $(\text{P} - \text{NoP})^{\text{Honest subjects}}$ does not contain fairness and payoff-related brain activation. In addition, the serial subtraction term above controls for any unspecific effects of personality

because the subjects in the ''promise possible'' condition have the same personality and display the same behavior as the subjects in the ''no promise possible'' condition. The above contrast thus rules out the impact of any personality differences on brain activation that have nothing to do with promise making and promise breaking.

Using the described serial subtraction terms, our study provides the opportunity to answer the following three research questions:

First, is it possible to differentiate between subjects who will break a promise and those subjects who will keep a promise based on the brain activity pattern during the promise stage of the paradigm, i.e., during a stage of the paradigm when the dishonest act might already be planned or prepared, but does not yet have to be implemented? In other words, can we predict whether subjects will keep or break the promise based on a perfidious brain activity pattern measured during the promise stage? We hypothesize that if subjects indeed already plan to break the promise at this stage of the paradigm, the misleading promise decision should evoke an emotional conflict. Such an emotional conflict might be indicated in the brain by increased activity in brain regions known to be involved in conflict (Baumgartner et al., 2008a; Botvinick et al., 1999) and in negative emotion processing (Amaral, 2003; Phillips et al., 2003; Sanfey et al., 2003), including anterior cingulate cortex, anterior insular cortex, or amygdala.

Second, there is another stage in the paradigm which takes place before subjects have to implement whether to keep or break their promise. Subjects receive the information during this stage that their investor is now deciding whether to trust or not. While the chosen promise level can positively affect the investor's trust decision in trust game trials with promise stage, this is not the case in trust game trials without promise stage. The investor's actual behavior is thus much more difficult to forecast in trust game trials without promise stage, and the negative outcome for the subjects (i.e., mistrust on the part of the investor) is more likely, making the anticipation process more uncertain and stressful. We therefore wondered whether this uncertain and stressful anticipation process might be more pronounced in subjects who intend to break rather than keep the promise. In other words, can we even differentiate between dishonest and honest subjects in a stage of our paradigm when no decision at all must be made? Recent brain imaging studies have consistently shown that the anticipation of such stressful and in particular uncertain events, that is events which can either be positive or negative, is primarily associated with increased activity in two brain regions, the bilateral anterior insula and right inferior frontal gyrus (Herwig et al., 2007a, 2007b). If it is indeed the case that this uncertain and stressful anticipation process were more pronounced in subjects who plan to break the promise, we would expect brain activation in the regions mentioned above.

Third, what are the differences in brain activity between breaking and keeping a promise when subjects must ultimately implement their decision? Previous studies on deception (for recent reviews see Sip et al., 2008; Spence et al., 2004) did not distinguish between the promise, anticipation, and the decision stage and focused instead on the act of implementing a lie. We argue that such a deceptive act involves a similar cognitive and emotional process as during the implementation of a broken promise. While deceptive subjects have to suppress the truthful response, dishonest subjects have to suppress the honest response. Either suppression most likely leads to an emotional conflict, which might include a guilty conscience or the fear of negative consequences in case the deceptive or dishonest act is detected. Deception paradigms have consistently associated this kind of conflictuous cognitive and emotional processes with increased activity of discrete anterior frontal regions and the anterior cingulate cortex (ACC). In addition, more recent studies, which increased the subjects' emotional involvement by using more ecologically valid paradigms (e.g., mock-crime scenarios, guilty knowledge tests; Abe et al., 2007; Kozel et al., 2005; Langleben et al., 2005) rather consistently showed increased activity in emotion-related areas, such as the amygdala, insula, and orbitofrontal cortex. Due to the similar cognitive and emotional processes assumed to take place in the promise breaker's brain, we expect a similar activity pattern in the decision stage of our paradigm in the contrast between subjects who break and those who keep a promise. However, it is important to note that our paradigm has two major advantages compared to previous deception paradigms (see Sip et al., 2008 for an extensive discussion of the limitations of previous deception paradigms), allowing us to study the mentioned processes in a more ecologically valid situation. First, while subjects in our paradigm were completely free to decide whether to break or keep the promise, subjects in all previous deception paradigms were forced to lie or to tell the truth. Second, while the dishonest act in our paradigm was embedded in a social exchange involving positive and negative consequences or costs for the exchange partners, the deceptive act in all previous deception studies did not have such consequences because the subjects were, without exception, interacting with the experimenter(s) (for the most ''realistic'' version, see Abe et al., 2007). Thus, in previous studies it was rather obvious to a subject that a lie could not cause any real harm or costs to the experimenter. However, lying without malevolent intent and without evoked consequences for the deceived individual lacks important elements of guilt, personal gain, and the psychological stress that often accompany the generation and enactment of a lie in the ''real world'' (Gneezy, 2005). For these reasons, our study is the first to explore the neural underpinnings of the emotional and cognitive processes discussed above using an ecologically valid paradigm where subjects could decide freely to break or keep the promise during a realistic social exchange involving positive or negative consequences for the exchange partners.

Summing up, our paradigm enables us to answer the following questions: Do subjects who ultimately breach or keep a promise already have a differential brain activation pattern in stages of the paradigm during which the decision to break or keep the promise does not yet have to be implemented, but might already be prepared or planned? In other words, can we predict the dishonest act based on perfidious brain activity in the promise or anticipation stage of the paradigm? Moreover, do we find a similar differential brain activation pattern during the decision stage of our paradigm

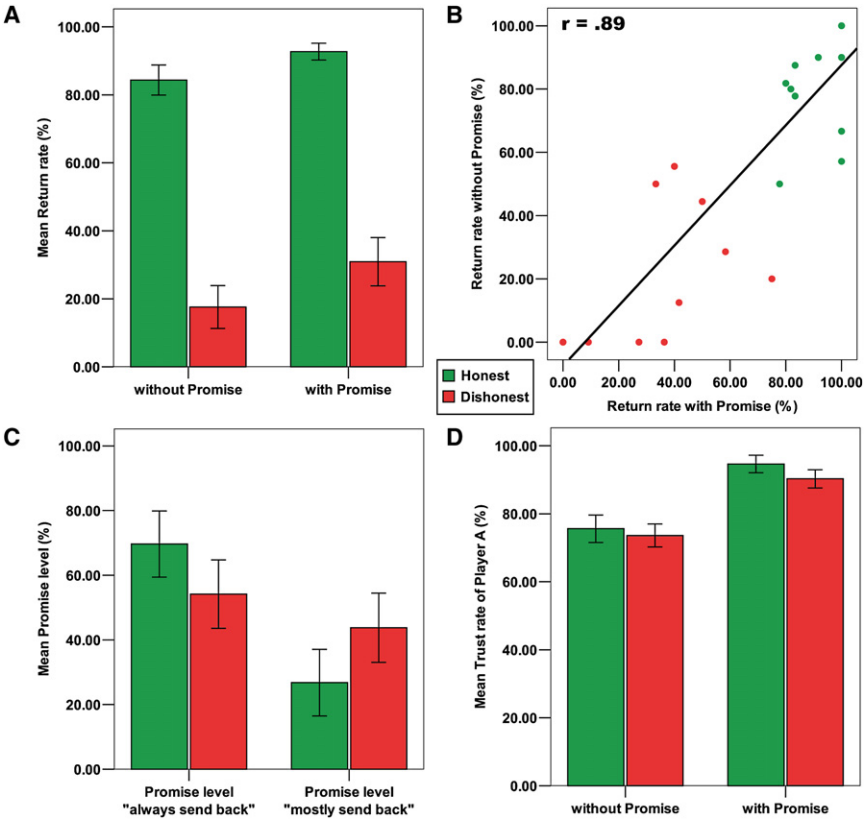

Figure 3. Behavioral Results

(A) Depicted are means ± SE of player B's return rates (in percentage), broken down for groups (dishonest/honest) and promise stages (trust games with/without antecedent promise stage). Findings indicate strong group differences in return rates irrespective of whether trust games are played with or without antecedent promise stage.

(B) High positive correlation (r = 0.89, p = 0.000) between return rates of trust games played with and without antecedent promise stage.

(C) Depicted are means ± SE of player B's promise levels (in percentage), broken down for groups (dishonest/honest) and the two highest promise levels (always send back/mostly send back). Findings indicate that both groups of subjects predominantly chose very high promise levels despite very different return rate patterns.

(D) Depicted are means ± SE of player A's trust rates (in percentage), broken down for groups (dishonest/honest) and promise stage (trust games with/without antecedent promise stage). Findings indicate no group differences in trust rates, but an increased trust rate, as expected, during trust game trials with antecedent promise stage.

between subjects who break and keep the promise as in the discussed deception studies where subjects were forced to lie or to tell the truth and where lying had no negative consequences for the deceived individual?

RESULTS

Behavioral and Psychometrical Results

Group Classification

Due to the fact that the trustees in our experiment were completely free to break or keep the promise, we examined in a first analysis whether our subjects can be classified into different subgroups based on their individual average return rate (see Experimental Procedures for details) in trust games played either with or without antecedent promise stage. For that purpose, we conducted a hierarchical cluster analysis (Ward's method, squared Euclidean distance measures, see Experimental Procedures for details) using both return rates (with and without antecedent promise stage) as dependent variables. Results indicated a cluster solution with two strongly separated clusters (see dendrogram of Figure S1). Inspection of the two clusters revealed two groups of subjects, i.e., those who either behaved trustworthily (referred to in the paper as *honest group/subjects*) or those who acted untrustworthily (referred to in the paper as *dishonest group/subjects*), irrespective of whether the trust games were played with or without antecedent promise stage (see Figure 3A). A two-way repeated-measures ANOVA with between-subject factor *group* (honest/dishonest) and within-subject factor *promise stage* (trust games with/without antecedent promise stage) revealed a highly significant main effect of *group* ($F_{(1,24)}$ = 102.80, p = 0.000, ETA^2 = 0.93), but no interaction effect of *group* × *promise stage* ($F_{(1,24)}$ = 0.46, p = 0.501, ETA^2 = 0.01), thus confirming that these two groups of subjects strongly differed in their return rate patterns, irrespective of whether the trust games were played with or without antecedent promise stage—a necessary precondition for the unconfounded analysis of the brain data as extensively discussed in the introduction section. The additionally discovered main effect of *promise stage* ($F_{(1,24)}$ = 8.86, p = 0.007, ETA^2 = 0.27) demonstrated that both groups of subjects showed some slight tendencies for increased return rates in trust game trials with antecedent promise stage (Figure 3A). Finally, the very high positive correlation between the two return rates (r = .89, p = 0.000, ETA^2 = 0.79) demonstrated not only that the two groups showed a consistent behavioral pattern but, importantly, that each individual subject alone did so as well (Figure 3B).

Promise Level

In a next analysis, we examined whether the two groups of subjects differed in their chosen promise level. The two lowest promise levels (*sometimes* or *never* send back half of the MUs) were only chosen three times in total (by three different subjects). Thus, subjects of each group chose one of the two highest promise levels during almost every promise decision, i.e., either always or mostly send back half of the MUs. Figure 3C illustrates the average of the two chosen highest promise levels (in percentage), broken down for the *dishonest* and *honest* group, respectively. A two-way repeated-measures ANOVA with between-subject factor *group* (dishonest/honest) and

within-subject factor *promise level* (always/mostly send back MUs) revealed neither main effects (main effect of group: $F_{(1,24)} = 0.209$, $p = 0.652$, $ETA^2 = 0.01$; main effect of promise level ($F_{(1,24)} = 3.264$, $p = 0.08$, $ETA^2 = 0.12$), nor an interaction effect (group × promise level: $F_{(1,24)} = 1.210$, $p = 0.282$, $ETA^2 = 0.05$), demonstrating that the two groups of subjects do not differ with respect to the chosen promise levels. Thus, the selection of different promise levels cannot explain the highly differential return rate pattern between the two groups during trust game trials with antecedent promise stage.

Trust Rate Player A

We next examined whether the differential return rates of player B are due to different trust rates of player A. We again calculated a two-way repeated-measures ANOVA with between-subject factor *group* (dishonest/honest) and within-subject factor *promise stage* (trust games with and without antecedent promise stage). Results revealed neither a main effect of *group* ($F_{(1,24)} = 0.957$, $p = 0.338$, $ETA^2 = 0.04$) nor an interaction effect of *group* × *promise stage* ($F_{(1,24)} = 0.131$, $p = 0.721$, $ETA^2 = 0.005$), suggesting that two groups experienced very similar trust rates of player A (Figure 3D). On the other hand, the main effect of promise stage was significant ($F_{(1,24)} = 29.408$, $p = 0.000$, $ETA^2 = 0.55$), demonstrating, as expected, an increased trusting behavior of player A in trust game trials with promise stage (Figure 3D).

Response Times

Next, we examined Player B's response times during both the promise and decision stages (excluding those trials during which Player B could not make a decision because player A did not trust him) using a two-way repeated-measures ANOVA with between-subject factor *group* (dishonest/honest) and within-subject factor *promise stage* (trust games with and without antecedent promise stage). We found no effect of group on response times (main effects of group and interaction effects of group × promise stage: all $p > 0.360$). The main effect of the factor *promise stage* during the decision trial was also not significant ($p > 0.254$), but, as expected, this main effect was significant during the promise stage ($F_{(1,24)} = 17.369$, $p = 0.000$, $ETA^2 = 0.42$), indicating an increase in response times during promise stages in which subjects actually had to decide about their promise level (mean ± SE: 3.14 s ± 0.19) compared to the other condition during which they just had to press a button without reflecting about the promise level (mean ± SE: 2.33 s ± 0.18; see Supplemental Experimental Procedures for details).

Personality Characteristics and Degree of Psychological Symptoms

Finally, we checked whether our two groups of subjects differ in main personality characteristics (e.g., neuroticism, extraversion, Machiavellism) and degree of psychological symptoms (e.g., depression, anxiety, aggression/hostility). For that purpose, we administered the "Brief Symptom Inventory" (BSI) questionnaire, the "NEO-Five-Factor-Inventory" (NEO-FFI) questionnaire (Costa and McCrae, 1992) and the Machiavelli questionnaire (Christie and Geis, 1970). Importantly, all scales showed no group differences (BSI: all $p > 0.33$, NEO-FFI: all $p > 0.30$, Machiavelli questionnaire: all $p > 0.21$). Furthermore, correlations of return rates with these personality and psychological symptom scales did not reveal any significant result (BSI: all $p > 0.38$, NEO-FFI: all $p > 0.22$, Machiavelli questionnaire: all $p > 0.29$; please see Tables S5–S7 for detailed statistical information to each scale). These findings suggest that the reported differential brain activity patterns (see below) are not driven by specific (related to the act of promising) personality differences between promise breakers and promise keepers, but that they rather reflect the (intended or actual act of) breaking a promise relative to the (intended or actual act of) keeping a promise, regardless of the subjects' personality characteristics. However, please note that the questionnaire evidence cannot completely rule out that an unknown personality or demographic factor not directly assessed by the questionnaires could contribute to the difference in the subjects' tendencies toward promise keeping or breaking.

Brain Imaging Results

Promise Stage

In a first brain imaging analysis, we were interested whether it is possible to differentiate between honest and dishonest subjects based on their brain activation pattern in the promise stage. This stage is of particular interest because, as we show in the behavioral results section, the two groups of subjects do not differ in their behavior, i.e., they chose the same promise level and even need the same amount of time to implement their decision. Furthermore, the promise stage takes place at a time point when the decision to be dishonest or honest does not yet have to be implemented, thus still providing the opportunity to reconsider and change the decision. It is therefore an open question whether subjects already show a perfidious brain activation pattern indicating the planned breach of promise at this time point. Comparing dishonest subjects with honest subjects (using the serial subtraction term: $[\text{Promise} - \text{No Promise}]^{\text{Dishonest subjects}} - [\text{Promise} - \text{No Promise}]^{\text{Honest subjects}}$) indeed revealed a highly differential brain activation pattern, i.e., dishonest subjects compared to honest subjects showed increased activation in the anterior cingulate cortex (ACC) and bilateral in the inferior frontal gyrus/anterior insula region (referred to as frontoinsular cortex in the following; Figures 4A and 4B, Table S1). In contrast, calculating the reversed serial subtraction term ($[\text{Promise} - \text{No Promise}]^{\text{Honest subjects}} - [\text{Promise} - \text{No Promise}]^{\text{Dishonest subjects}}$) showed no increased activation in honest compared to dishonest subjects, even at a strongly lowered $p < 0.05$ (uncorrected).

In order to clarify whether the revealed brain activation pattern is not only group, but also stage-specific, we created functional regions of interests (see Supplemental Experimental Procedures for details) in the ACC and bilateral frontoinsular cortex and extracted, based on these ROIs, β estimates in all stages of the paradigm, including the anticipation and decision stages (decision phase A + B). We calculated independent t tests based on these β estimates in order to check for group differences in these brain regions. We found no other stage of the paradigm in which these regions showed a differential group effect (ACC: all $p > 0.29$; right frontoinsular cortex: all $p > 0.26$; left frontoinsular cortex: all $p > 0.42$), indicating that this neural correlate is both group-dependent and stage-dependent; that is, only subjects of the dishonest group who later intend to break their promises in the decision stage react with increased activation in the ACC and bilateral frontoinsular cortex during the promise stage.

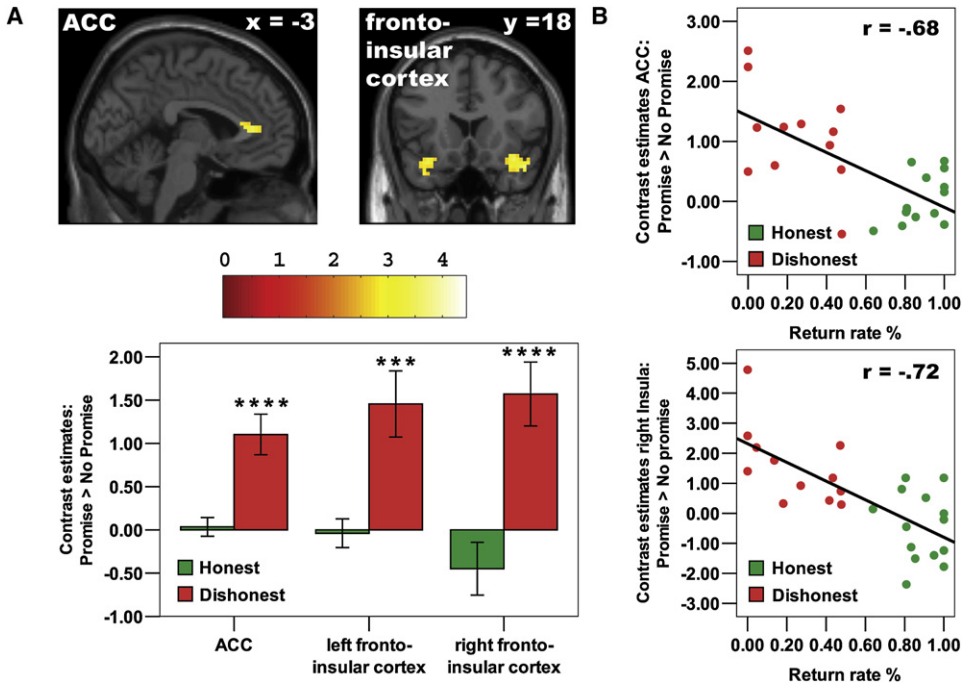

Figure 4. Differential Brain Activation Pattern during the Promise Stage

(A) Depicted on sagittal and coronal slices is the increased activation in dishonest compared to honest subjects (based on the serial subtraction term: [*Promise – No Promise*]$^{Dishonest\ subjects}$ – [*Promise – No Promise*]$^{Honest\ subjects}$) in the ACC (BA 24, x = −6, y = 33, z = 6) and bilateral frontoinsular cortex (BA 47/13, x = −30, y = 24, z = −18; x = 42, y = 15, z = −24) at $p < 0.005$ (voxel extent threshold: 10 voxels, for display purposes depicted at $p < 0.01$). Despite the fact that both groups of subjects implement the same promise decision, the dishonest subjects who will deceive at the following decision stages already show a perfidious brain activation pattern during the promise stage. Bar plots representing contrast estimates ± SE (*Promise > No Promise*) of functional ROIs (see Experimental Procedures for details) demonstrate that the differential group effect in all regions is mainly based on increased activation in the dishonest group in the *Promise* compared to the *No Promise* condition at $p \leq 0.005$ (***) or $p \leq 0.001$ (****).

(B) Return rates show a strong negative correlation with ACC ($r = –0.68$, $p < 0.001$) and bilateral frontoinsular cortex (right frontoinsular cortex: $r = –0.72$, $p < 0.001$; left frontoinsular cortex [not depicted]: $r = –0.66$, $p < 0.001$) using the same functional ROIs as in (A).

Anticipation Stage

In a next analysis, we were interested whether dishonest and honest subjects also show differential brain activations in the anticipation stage of the paradigm, that is in a stage of the paradigm during which no decision related the dishonest or honest act has to be made. We focused in our analysis in particular on the anticipation process during trust game trials *without* antecedent promise stage. In these trials, in contrast to trials *with* antecedent promise stage, choosing a high promise level cannot influence the investor's actual behavior, making the anticipation process more uncertain and stressful. We indeed found that the two groups differ in this uncertain and stressful anticipation process. Comparing dishonest with honest subjects (using the serial subtraction term: [No Promise - Promise]$^{Dishonest\ subjects}$ minus [No Promise – Promise]$^{Honest\ subjects}$) revealed increased brain activation in the right anterior insula and right inferior frontal gyrus (IFG) in dishonest subjects (Figures 5A and 5B, Table S2). In contrast, calculating the reversed serial subtraction term showed no increased activation in honest compared to dishonest subjects, even at a strongly lowered $p < 0.05$ (uncorrected), suggesting that this anticipation process is more pronounced in subjects who behave dishonestly.

In a next step, we again examined how stage-specific this activation pattern actually is. For that purpose, we extracted β estimates based on functional ROIs (IFG and anterior insula) for all stages of the paradigm. Independent t tests revealed no differential group effect in these brain regions during any other stage of the paradigm (IFG: all $p > 0.19$; anterior insula: all $p > 0.44$), again indicating that the activation in these brain regions is not only group, but also stage dependent.

Decision Stage

We used two different regression models to examine the brain activation pattern during the decision stage. In a first model of the decision stage, we were interested in brain regions showing a sustained activation over both decision phases A + B (decision phase A, revealment of player A's trust decision; decision phase B, player B is reminded of his promise, see Figure 2 for a detailed explanation of these two phases). For that purpose, we created a decision regressor which modeled the decision epoch as a whole, i.e., from onset decision screen in decision phase A until implementation of the decision via button press in decision phase B (mean duration 10.13 s). In a second model of the decision stage, we modeled decision phases A and B separately, in order to examine whether the two phases can be differentiated

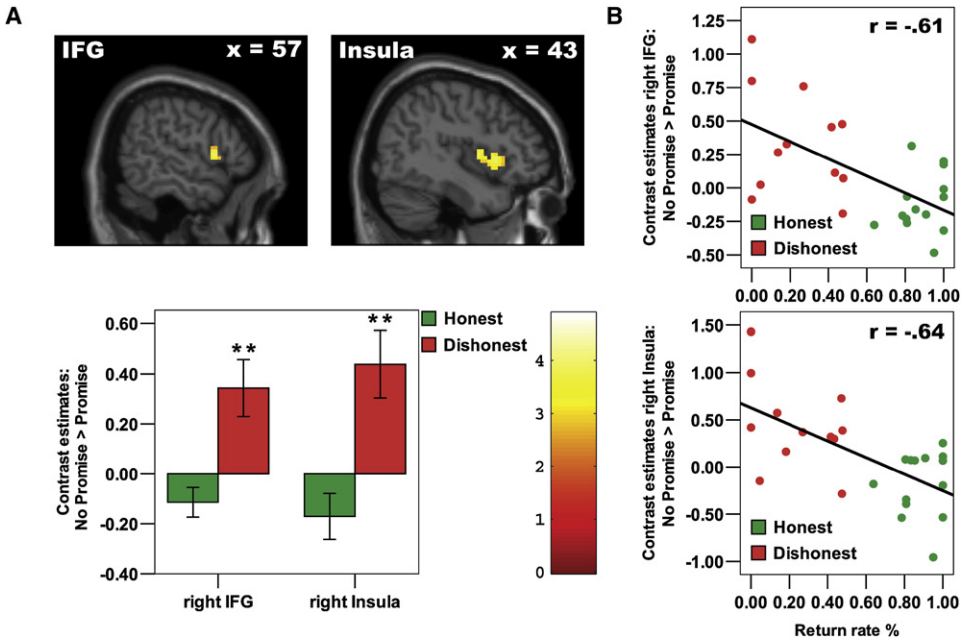

Figure 5. Differential Brain Activation Pattern during the Anticipation Stage

(A) Depicted on sagittal slices is the increased activation in dishonest compared to honest subjects (based on the serial subtraction term: [*No Promise* – *Promise*]$^{Dishonest\ subjects}$ – [*No Promise* – *Promise*]$^{Honest\ subjects}$) in the right IFG (BA 45, x = 57, y = 12, z = 6) and right anterior insula (BA 13, x = 45, y = 0, z = 6) at $p < 0.005$ (voxel extent threshold: 10 voxels, for display purposes depicted at $p < 0.01$). Despite the fact that both groups are confronted with the same uncertainty during the anticipation of player's A trusting behavior (whether or not he trusts), the brain activation pattern of the dishonest subjects suggests a more pronounced anticipation process. Bar plots representing contrast estimates ± SE (*No Promise* > *Promise*) of functional ROIs (see Experimental Procedures for details) demonstrate that the differential group effect in all regions is mainly based on increased activation in the dishonest subjects in the *No Promise* compared to the *Promise* condition at $p \leq 0.01$ (**).

(B) Return rates show a strong negative correlation with right IFG ($r = –0.61$, $p < 0.001$) and right anterior insula ($r = –0.64$, $p < 0.001$) using the same functional ROIs as in (A).

by a unique brain activation pattern (for details of the two different models please see Supplemental Experimental Procedures).

Examining the decision stage as whole (using the decision regressor of the first model) by comparing the dishonest subjects with the honest subjects (using the serial subtraction term: [Promise – No Promise]$^{Dishonest\ subjects}$ – [Promise – No Promise]$^{Honest\ subjects}$) revealed only one brain region that showed a differential activity: the dishonest subjects showed sustained activation in the ventral part of the striatum during the whole decision stage (Figure 6, Table S3). In contrast, a separate examination of the two decision phases (based on the decision regressors of the second model) using the same serial subtraction term revealed increased activation in dishonest subjects in the ACC and left DLPFC (at the border between DLPFC and VLPFC) during decision phase A (Figures 7A and 7B, Table S3), while the same group of subjects showed increased activation in the left amygdala during decision phase B (Figure 7C, Table S3). We observed no increased brain activation using the reversed serial subtraction terms in honest compared to dishonest subjects, even at a strongly lowered $p < 0.05$ (uncorrected).

In order to corroborate the described specificity in the decision stage, we created functional ROIs and extracted β estimates separately for all three decision regressors (decision phase A + B, decision phase A, and decision phase B). Independent t tests confirmed the suggested specificity with respect to the time point of differential group activity during the decision stage for all ROIs (ventral striatum, DLPFC, ACC, and amygdala; please see Table S4 for details). Independent t tests of β estimates based on the same functional ROIs of the decision stage also showed no differential group effect during any other stage (promise and anticipation) of the paradigm (ACC: all $p > 0.08$; DLPFC: all $p > 0.32$; amygdala: all $p > 0.45$; ventral striatum: all $p > 0.93$).

Finally, we conducted additional analyses presented in the supplementary material in order to further control for potential confounding factors (Supplemental Analysis S1), to further corroborate the stage-specificity of the activity patterns (Supplemental Analysis S2), and to examine the activity in the decision stage with slightly different decision regressors (Supplemental Analysis S3). These three additional analyses confirmed the findings reported above.

DISCUSSION

In order to study the neural underpinnings of nonbinding cooperative agreements in the form of promises, we used a social-interaction paradigm derived from game theory in which subjects were completely free to decide whether to break or to keep the promise and in which breaking or keeping a promise

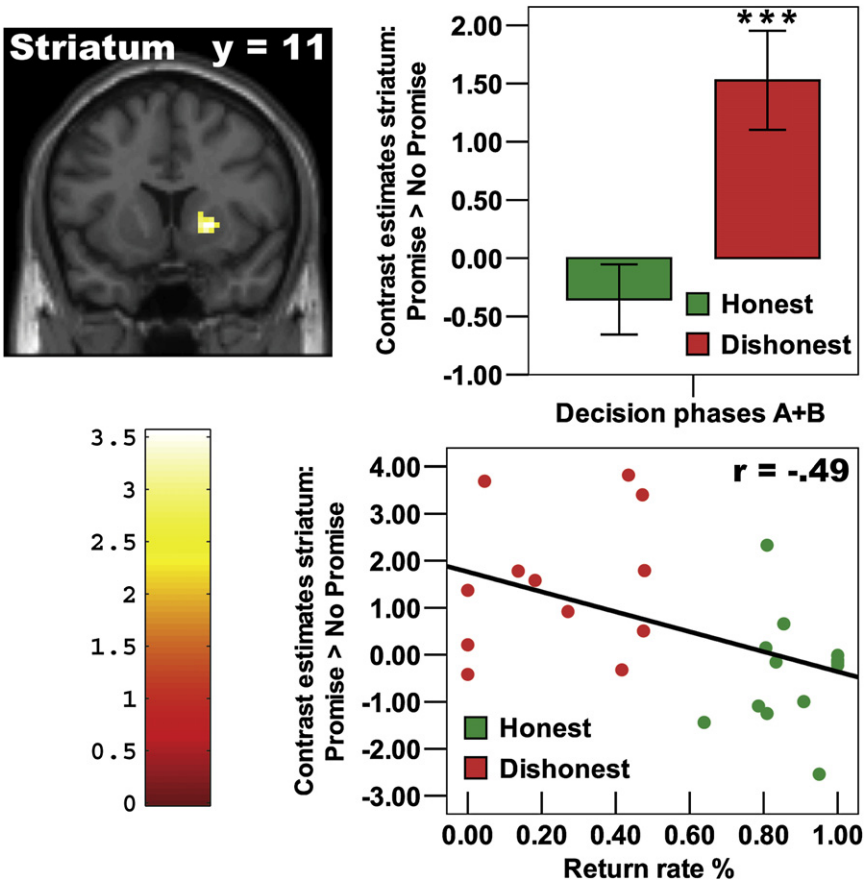

Figure 6. Differential Brain Activation Pattern during the Decision Stage with Combined Modeled Decision Phases A and B

Depicted on a coronal slice is the increased activation in dishonest compared to honest subjects (based on the serial subtraction term: [*Promise – No Promise*]$^{Dishonest\ subjects}$ – [*Promise – No Promise*]$^{Honest\ subjects}$) in the right ventral striatum ($x = 24$, $y = 12$, $z = 0$) at $p < 0.005$ (voxel extent threshold: 10 voxels, for display purposes depicted at $p < 0.01$). This finding suggests that dishonest subjects have increased activity in the ventral striatum during the whole decision process. Bar plots representing contrast estimates ± SE (*Promise* > *No Promise*) of functional ROIs (see Experimental Procedures for details) demonstrate that the differential group effect is mainly based on increased activation in dishonest subjects in the *Promise* compared to the *No Promise* condition at $p \leq 0.005$ (***). The scatter plot demonstrates that the return rates are negatively correlated with activity in the right ventral striatum ($r = -0.49$, $p < 0.01$) using the same functional ROI.

caused monetary consequences (benefits or costs) for both exchange partners. We found that all stages of the paradigm revealed a highly specific brain activation pattern, enabling us to differentiate between subjects who break a promise and those who keep a promise (see Figures 4–7). Importantly, the applied serial subtraction term analysis (see Introduction and Results section) rules out the impact of any personality differences on brain activation that have nothing to do with promise making and promise breaking. Furthermore, the obtained questionnaire evidence favors the view that the reported differential brain activity patterns are also not driven by specific (related to the act of promising) personality differences between promise breakers and promise keepers, but rather that they reflect the (intended or actual) act of breaking a promise relative to the (intended or actual) act of keeping a promise, regardless of the subjects' personality characteristics. However, please note that the questionnaire evidence does not completely rule out that other unknown personality factors, which are not directly assessed by the questionnaires, contribute to the difference in the subjects' tendencies toward promise keeping or breaking.

Two stages of the paradigm allow us to look for differences in brain activity between honest and dishonest subjects during time points when the subjects do not yet have to implement the decision to break or to keep the promise. The stage of particular interest in this regard is the promise stage of the paradigm because behavioral findings in our study showed that dishonest and honest subjects do not differ with regard to their chosen promise level, and even the response times for implementing the promise decision are equal. Nevertheless, the brain activation pattern is highly differential, that is subjects who will break their promise at later stages of the paradigm already show increased activation in the ACC and bilateral frontoinsular cortex. The ACC has been demonstrated to be consistently implicated in conflict monitoring and cognitive control both during social (Baumgartner et al., 2008a; Delgado et al., 2005) and nonsocial paradigms (Botvinick et al., 1999, 2001; Carter et al., 1998). The insula (including frontoinsular cortex) has been shown to be involved in the mapping of body-related sensations, including temperature, pain, proprioception, and viscera (for review see Craig, 2002). Consistent with this mapping hypothesis, insula activations were mainly found during aversive emotional experiences associated with strong visceral and somatic sensations such as the experience of unfairness (Sanfey et al., 2003; Tabibnia et al., 2008; Tabibnia et al., 2008), the threat of punishment (Spitzer et al., 2007), and the anticipation of negative and unknown emotional events (Herwig et al., 2007a, 2007b). Taken together, the increased activation in the ACC and bilateral frontoinsular cortex suggests that subjects who behave dishonestly already form their intent to break the promise during the promise stage. We assume that this intention leads to a decision conflict and associated (aversive) emotional experiences, represented in the brain in the ACC and frontoinsular cortex. The aversive emotional experience might include the guilty conscience toward the exchange partner whom the promise will intentionally mislead. Interestingly, both of these brain regions are thought to belong to a reflexive, automatic system of social cognition proposed

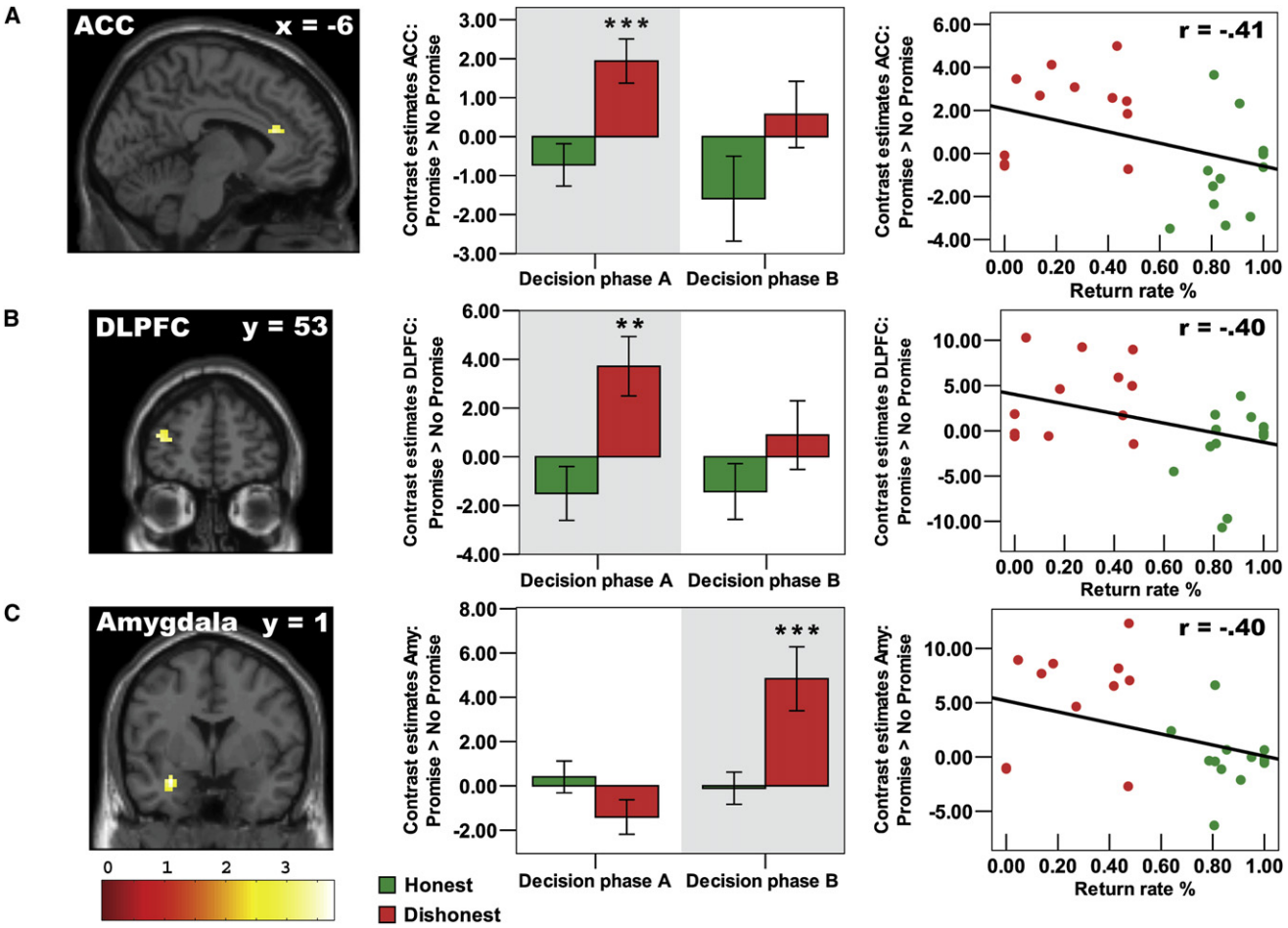

Figure 7. Differential Brain Activation Pattern during the Decision Stage with Separately Modeled Decision Phases A and B

Depicted on sagittal and coronal slices is the increased activation in dishonest compared to honest subjects (based on the serial subtraction term: [*Promise – NoPromise*]$^{Dishonest\ subjects}$ – [*Promise – No Promise*]$^{Honest\ subjects}$) during *decision phase A or B* at $p < 0.005$ (voxel extent threshold: 10 voxels, for display purposes depicted at $p < 0.01$). In *decision phase A*, increased activation was found in the (A) ACC (BA 24, x = −6, y = 27, z = 18) and (B) left DLPFC (BA 10/46, x = −39, y = 54, z = 15), whereas in *decision phase B* increased activity was found in the (C) left amygdala (x = −30, y = 0, z = −21). Bar plots representing contrast estimates ± SE (*Promise > No Promise*) of functional ROIs (see Experimental Procedures for details) confirm this suggested activity pattern by illustrating the group-dependent and phase-dependent activity of these brain regions during the two phases of the decision stage. Asterisks indicate significantly increased activity in dishonest subjects in the *Promise* compared to the *No Promise* condition at $p \leq 0.01$ (**) or $p \leq 0.005$ (***). Finally, the scatter plots demonstrates that the return rates are negatively correlated with activity in the ACC ($r = –0.41$, $p < 0.05$) and left DLPFC ($r = –0.40$, $p < 0.05$) during *decision phase A* as well as left amygdala ($r = –0.40$, $p < 0.05$) during *decision phase B*.

by Lieberman and colleagues (Lieberman, 2007; Satpute and Lieberman, 2006). We thus speculate that due to the reflexive mode of operation of these brain regions, it might be rather difficult or even impossible for dishonest subjects to suppress this reaction pattern in the brain voluntarily, i.e., not to "signal" their planned breach of promise with a perfidious brain activation pattern.

Another stage of the paradigm takes place before the dishonest act has to be implemented. During this stage, the subjects do not even have to make a decision, they are merely informed that their exchange partners are now deciding whether to trust or not and the subjects can thus do nothing but anticipate the outcome of the investor's trust decision. Interestingly, the two groups (dishonest/honest) do not differ in hypothesized regions of interests during anticipation trials *with* antecedent promise stage (see Table S2 for the two small differences in other regions). In these trials, choosing a high promise level can influence the investor's trusting behavior (and all subjects did so), thus reducing the probability that the investor will not trust. In contrast, the investor's trusting behavior cannot be affected in trials *without* antecedent promise stage and the outcome of the trust decision is therefore much more difficult to forecast, making the anticipation trial more emotional and stressful. Recent brain imaging studies (Herwig et al., 2007a, 2007b, 2009) have shown that the anticipation of such negative and unforeseeable (either negative or positive) emotional events is mainly associated with increased activation in the bilateral anterior insula and right IFG. Moreover, these studies show

that personality traits of depression and neuroticism, both of which are associated with negative expectations toward future events, correlate positively with these brain regions during the anticipation trials, i.e., the higher the score in these personality measures, the higher the activation in the bilateral anterior insula and right IFG. We found that similar to subjects with higher depressive or neuroticism scores, subjects who behaved dishonestly reacted to the unpredictable and thus emotional and stressful anticipation stage of our paradigm with increased activation in the same brain regions (right anterior insula and right IFG). This suggests that social exchange situations associated with a lack of control and uncertainty are more pronounced and more intensely experienced in subjects who intend to behave dishonestly, which might indicate that they more strongly anticipate a negative outcome (e.g., mistrust on the part of the investor) in unpredictable social situations than subjects who intend to behave honestly. Taken together, our findings demonstrate that the dishonest subjects can be differentiated from honest subjects even in stages of the paradigm during which no decision related to the dishonest act has to be made.

The stage during which the dishonest or honest act actually has to be implemented revealed an activity pattern in accordance with our assumption that the *breaking of a promise* and the *telling of a lie* involve similar cognitive and emotional processes and associated brain activation patterns. In detail, we argued that while deceptive subjects have to suppress the truthful response, dishonest subjects have to suppress the honest response. In line with this assumption, our study, along with most deception paradigms (e.g., Abe et al., 2006; Kozel et al., 2005; Lee et al., 2005; Nuñez et al., 2005; Phan et al., 2005; Spence et al., 2001, 2008), revealed increased activity in brain regions of the lateral PFC which are known to play an essential role in the control and suppression of (inappropriate) cognitions and behaviors (e.g., Aron, 2007; Baumgartner et al., 2006, 2008b; Beeli et al., 2008; Jäncke et al., 2008; Spitzer et al., 2007). Furthermore, we argued that the suppression of both the truthful and the honest response most likely leads to an emotional conflict in the deceptive and dishonest subjects. Again corroborating this assumption, our study and most of previous deception studies (e.g., Abe et al., 2006; Kozel et al., 2005; Langleben et al., 2005; Lee et al., 2005; Nuñez et al., 2005; Phan et al., 2005; Spence et al., 2001) demonstrated increased activity in the ACC, which constitutes the brain region most consistently associated with cognitive and emotional conflict processing and resolving (e.g., Baumgartner et al., 2008a; Botvinick et al., 1999; Etkin et al., 2006). Taken together, our paradigm, which substantially improved previous deception paradigms (subjects in our paradigm were free to decide and their decisions caused both positive and negative consequences for the exchange partners, see Introduction), confirmed the activation of the aforementioned brain regions during the assumed cognitive and emotional processes involved in the implementation of the deceptive or dishonest acts. Moreover, our paradigm also substantiated the assumption that truthful responding comprises a *relative* baseline in human cognition and communication (i.e., truthful responding compared to lying does not require an activity increase in any single brain region, Spence et al., 2004), because, similar to most deception paradigms, we did not find any activity increase during the decision stage of our paradigm in subjects who behaved honestly compared to those who behaved dishonestly. Furthermore, we could extend these negative findings to all other stages of our paradigm (promise and anticipation stages). Thus, in spite of the fact that our honest subjects freely chose to keep their promises in a "realistic" social exchange, no specific neural correlate of honesty was observed in any stage of the paradigm, even at a strongly lowered significance threshold.

Besides increased activity in the ACC and DLPFC, the amygdala demonstrated increased activity during the breaking of a promise in the decision stage of our paradigm. Whereas activity in the ACC and DLPFC belong to the most replicated findings in neuroimaging studies on deception, up to now only three of the deception studies reported increased activation of the amygdala—a brain region widely acknowledge to play an important role in emotion (Phan et al., 2002; Phillips et al., 2003) and in particular fear processing (Adolphs et al., 2005; Amaral, 2003; Baumgartner et al., 2008a). In two of these studies, subjects had to detect deceptive intentions; the findings indicated that the crucial factor for amygdala activation is the subject's involvement, that is, amygdala activation was only observed if the subject was the target of the deceit (Grèzes et al., 2004, 2006). Only one study, which focused on the neural activities of those telling lies, reported activation of the amygdala. Of all conducted deception studies, this study (Abe et al., 2007) used a paradigm that might get closest to real life deception by introducing a clever twist in the paradigm. This twist consisted of having a second experimenter tell the subject to disobey the first experimenter, i.e., when the first experimenter instructed the subject to tell the truth, the second experimenter secretly asked the subject to deceive. Thus, we conclude that increasing the subjects' emotional involvement by creating a "realistic" social situation seems to trigger the amygdala response in the study by Abe and colleagues (2007) and our paradigm—notably in a very similar ventral part of the left amygdala. Furthermore, the time point of amygdala activation in our paradigm provides some additional evidence as to which process might have evoked the amygdala activation in both studies. This evidence can be derived from the fact that we only found increased activation of the amygdala during decision phase B, i.e., when subjects were reminded of their promise they were going to break. This suggests that it is not the dishonest or deceptive act per se (including the inhibition of the honest/truthful response and associated conflict), but rather the deliberate confrontation with the promise toward the interaction partner, which might drive the amygdala activation. Whereas subjects in our paradigm explicitly had to make a promise toward the interaction partner, the promise was more implicit in nature in the study of Abe and colleagues (2007), i.e., subjects implicitly promised the first experimenter to obey his instructions. Taken together, we argue that the spontaneous (study of Abe et al.) or triggered (our paradigm) reminder of a promise one is not allowed (study of Abe et al.) or willing (our paradigm) to keep evokes an emotional response in deceptive or dishonest subject, which might include a guilty conscience toward the interaction partner and/or a fearful reaction that the deceptive or dishonest act will be detected.

Finally, we found increased activation in the right ventral striatum during the breaking of a promise in the decision stage of our paradigm. Similar to the observed activity of the amygdala, only very few of the discussed deception studies reported activations in the striatum (e.g., Nuñez et al., 2005); these activations are commonly observed during tasks that require individuals to suppress a prepotent or frequent response (Aron et al., 2007; Casey et al., 2002). Thus, the activity in the striatum may reflect, similar to the activation in the left DLPFC, the inhibition of the impulse to answer truthfully or honestly. However, we suggest an alternative interpretation for the striatum activation in our paradigm for the following reasons. First, in contrast to the DLPFC activation, which was restricted to phase A of the decision stage, we observed sustained activation in the ventral striatum during the entire time window of the decision stage, suggesting a different cognitive or affective process. Second, in contrast to the few deception studies which reported activation in this brain region, our study used a social exchange paradigm in which subjects deliberately decided to break the promise with the goal of increasing their monetary payoff at the expense of the exchange partner. Due to the well-known role of the striatum in social (e.g., Fliessbach et al., 2007; Rilling et al., 2004) and nonsocial (Delgado et al., 2004; Liu et al., 2007) reward processing and its strong impact on decision making (de Quervain et al., 2004; Delgado et al., 2005, 2008; for a recent review, Fehr and Camerer, 2007; King-Casas et al., 2005; Knutson et al., 2007), we thus speculate that the activation in the striatum might represent the motivational, appetitive component of the dishonest act. In other words, subjects might be motivated to break the promise because the activation in the ventral striatum reinforces the dishonest act and thus might act as a counterbalance against the aversive emotions (e.g., guilty conscience) and potential negative consequences in case the deception should be detected. We suggest designing future studies that allow examining whether the former, the latter, or both interpretations for the striatum activity apply.

Summing up, this study explored the neural correlate of nonbinding cooperative agreements in the form of a promise—one of the oldest human-specific psychological mechanisms fostering trust, cooperation, and partnership formation. In order to study this psychological mechanism, we applied a social interaction paradigm derived from game theory in which subjects were completely free to decide whether to keep or break the promise and in which the dishonest act included both benefits for the subjects and costs for the exchange partners. Findings revealed that each of the three processes playing an important role during nonbinding cooperative agreements is associated with a unique brain activation pattern, allowing us to discriminate dishonest from honest subjects. In detail, we found (1) that the implementation of the dishonest act is associated with increased activity in brain regions known to be involved in cognitive control and conflict processing, including the DLPFC and ACC. In addition, we also demonstrated (2) increased activation during this stage of the paradigm in emotion-related brain regions, including amygdala and ventral striatum. We suggest that the amygdala activation may represent the guilty conscience or the fear that the deceptive act could be detected, whereas the activity in the ventral striatum might represent the motivating and driving force behind the deceptive act. Finally, one of the most important findings concerns (3) the predictive power of ''perfidious'' brain activation patterns in the ACC, bilateral frontoinsular cortex, and right IFG during the promise or the anticipation stages for the final decision whether to keep or break the promise. Even though during the promise stage the *behavior* of those subjects who ultimately cheat their exchange partner and those who finally keep their promise does not differ—both types of subject promise to keep the informal agreement—the *brain activations* of the ''cheaters'' and the ''promise keepers'' show very distinct patterns during the promise stage. These findings contribute to a recent debate about whether data from neuroscience are relevant for sciences such as economics that are primarily interested in understanding and predicting behavior (Camerer et al., 2005; Glimcher and Rustichini, 2004). The fact that the cheaters' brain activations during the promise and anticipation stages differ unambiguously from those of the promise keepers, even though both of them perform the same behavior, means that the brain activations alone and not just the observed behaviors are capable of predicting the dishonest act. Thus, our study shows that data from neuroscience can provide important insights into behavior that extend beyond that which purely behavioral data can detect.

EXPERIMENTAL PROCEDURES

Subjects

A total of 34 healthy male students from different universities in Zurich participated in the study. Eight of the participants had to be excluded from the analyses; one subject due to scanner malfunctions and another seven subjects due to design constraints (see Supplemental Experimental Procedures for details), resulting in 26 male subjects (mean age ± SD, 23.5 ± 2.5) for the analyses of the behavioral and brain imaging data. All subjects were free of chronic diseases, mental disorders, medication, and drug or alcohol abuse. The study was carried out in accordance with the Declaration of Helsinki principles and approved by the institutional ethics committee. All subjects gave written, informed consent and were informed of their right to discontinue participation at any time. Subjects received a lump sum payment of CHF 40 for participating in the experiment plus the additional money earned during the trust game trials (exchange rate 10 money units = 2.5 Swiss Franc, that is about $2.50).

Design

In total, subjects played 24 trust game trials in the role of a trustee (player B) against 24 different and anonymous human interaction partners in the role of an investor (player A, see Figures 1 and 2 and Supplemental Experimental Procedures for details). In half of these trials, subjects had to make a promise for three subsequently played trust game trials whether they *always*, *mostly*, *sometimes*, or *never* plan to send back half of the money so that both players earn the same amount. Importantly, player A was always informed about B's promise, and B could keep the promise, but he was also allowed to break it. In total, player B made four promise decisions and each of these decisions held for the three subsequent trust game trials. There were also four instances during which player B was informed that he could not decide on a promise level; the three succeeding trust game trials were thus played without promise. Trust game trials with and without antecedent promise stage were presented counterbalanced and pseudorandomized.

Behavioral Analysis

We created two return rate indexes for the behavioral data (return decisions)—one for trust game trials with antecedent promise stage and one for trials without antecedent promise stage. The index measures player B's average return rate for the trust game trials in which player A trusted, i.e., the

percentage of cases in which B proved trustworthy and equalized payoffs. Using these two behavioral indexes, we performed a hierarchical cluster analysis based on the Ward method (using the squared Euclidean distance measure) in order to classify our subjects into different subgroups. This cluster analysis revealed a cluster solution with two strongly separated clusters (see dendrogram of Figure S1). Inspection of the two clusters revealed two groups of subjects, i.e., those who either behaved trustworthily (referred to in the paper as *honest group/subjects*) or those who acted untrustworthily (referred to in the paper as *dishonest group/subjects*). Please see Supplemental Experimental Procedures section for further information on the analyses of the behavioral data, including promise levels, response times and trust rates of player A.

fMRI Acquisition

The experiment was conducted on a 3 Tesla Philips Intera whole-body MR Scanner (Philips Medical Systems, Best, The Netherlands) equipped with an eight-channel Philips SENSE head coil. Structural image acquisition consisted of 180 T1-weighted transversal images (0.75 mm slice thickness). For functional imaging, a total of 380 volumes were obtained using a SENSitivity Encoded (SENSE; Pruessmann et al., 1999) T2*-weighted echo-planar imaging sequence with an acceleration factor of 2.0. Forty-two axial slices were acquired covering the whole brain with a slice thickness of 3 mm; no interslice gap; interleaved acquisition; TR = 3000 ms; TE = 35 ms; flip angle = 77°, field of view = 220 mm; matrix size = 80 × 80. We used a tilted acquisition in an oblique orientation at 30° to the AC-PC line in order to optimize functional sensitivity in orbitofrontal cortex and medial temporal lobes.

fMRI Analysis

Data were preprocessed and statistically analyzed using SPM5. For preprocessing, all images were realigned to the first volume, corrected for motion artifacts and time of acquisition within a TR, normalized into standard stereotaxic space (template provided by the Montreal Neurological Institute), and smoothed using an 8 mm full-width-at-half-maximum Gaussian kernel. For statistical analysis, we performed random-effects analyses on the functional data for the promise, anticipation, and decision stage. For that purpose, we estimated two general linear models (GLMs) and computed linear contrasts of regression coefficients at the individual subject level. In order to enable inference at the group level, we calculated second-level group contrasts using independent t tests with factor group (dishonest/honest group), separately for each stage of the paradigm. We applied an uncorrected p value of 0.005 combined with a cluster-size threshold of 10 voxels to our apriori regions of interests (see Introduction). Furthermore, we checked whether our a priori regions of interests survive small volume family-wise-error (FWE) corrections at $p < 0.05$. Crucially, all our regions of interests survived this correction procedure. Please see Supplemental Experimental Procedures for additional information on all conducted statistical analyses, including a more detailed description of the applied GLMs and multiple comparison corrections.

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures, four tables of brain activity, three tables of questionnaire measures, one figure of the cluster analysis (Dendrogram) and three analyses of brain activity and can be found with this article online at http://www.cell.com/neuron/supplemental/S0896-6273(09)00900-3.

ACKNOWLEDGMENTS

This work is part of Project 9 of the National Competence Center for Research (NCCR) in Affective Sciences. The NCCR is financed by the Swiss National Science Foundation. E.F. also gratefully acknowledges support from the research priority program at the University of Zurich on the "Foundations of Human Social Behavior."

Accepted: November 5, 2009
Published: December 9, 2009

REFERENCES

Abe, N., Suzuki, M., Tsukiura, T., Mori, E., Yamaguchi, K., Itoh, M., and Fujii, T. (2006). Dissociable roles of prefrontal and anterior cingulate cortices in deception. Cereb. Cortex *16*, 192–199.

Abe, N., Suzuki, M., Mori, E., Itoh, M., and Fujii, T. (2007). Deceiving others: distinct neural responses of the prefrontal cortex and amygdala in simple fabrication and deception with social interactions. J. Cogn. Neurosci. *19*, 287–295.

Adolphs, R., Gosselin, F., Buchanan, T.W., Tranel, D., Schyns, P., and Damasio, A.R. (2005). A mechanism for impaired fear recognition after amygdala damage. Nature *433*, 68–72.

Amaral, D.G. (2003). The amygdala, social behavior, and danger detection. Ann. N.Y. Acad. Sci. *1000*, 337–347.

Aron, A.R. (2007). The neural basis of inhibition in cognitive control. Neuroscientist *13*, 214–228.

Aron, A.R., Durston, S., Eagle, D.M., Logan, G.D., Stinear, C.M., and Stuphorn, V. (2007). Converging evidence for a fronto-basal-ganglia network for inhibitory control of action and cognition. J. Neurosci. *27*, 11860–11864.

Baumgartner, T., Valko, L., Esslen, M., and Jäncke, L. (2006). Neural correlate of spatial presence in an arousing and noninteractive virtual reality: an EEG and psychophysiology study. Cyberpsychol. Behav. *9*, 30–45.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., and Fehr, E. (2008a). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. Neuron *58*, 639–650.

Baumgartner, T., Speck, D., Wettstein, D., Masnari, O., Beeli, G., and Jancke, L. (2008b). Feeling present in arousing virtual reality worlds: prefrontal brain regions differentially orchestrate presence experience in adults and children. Frontiers in Human Neuroscience *2*, 8.

Beeli, G., Casutt, G., Baumgartner, T., and Jäncke, L. (2008). Modulating presence and impulsiveness by external stimulation of the brain. Behav. Brain Funct. *4*, 33.

Behrens, T.E., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F. (2008). Associative learning of social value. Nature *456*, 245–249.

Botvinick, M., Nystrom, L.E., Fissell, K., Carter, C.S., and Cohen, J.D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature *402*, 179–181.

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., and Cohen, J.D. (2001). Conflict monitoring and cognitive control. Psychol. Rev. *108*, 624–652.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., and Marois, R. (2008). The neural correlates of third-party punishment. Neuron *60*, 930–940.

Camerer, C., Loewenstein, G., and Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. J. Econ. Lit. *43*, 9–64.

Carter, C.S., Braver, T.S., Barch, D.M., Botvinick, M.M., Noll, D., and Cohen, J.D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. Science *280*, 747–749.

Casey, B.J., Thomas, K.M., Davidson, M.C., Kunz, K., and Franzen, P.L. (2002). Dissociating striatal and hippocampal function developmentally with a stimulus-response compatibility task. J. Neurosci. *22*, 8647–8652.

Charness, G., and Dufwenberg, M. (2006). Promises and partnership. Econometrica *74*, 1579–1601.

Christie, R., and Geis, F. (1970). Studies in Machiavellism (New York: Academic Press).

Costa, P.T., and McCrae, R.R. (1992). Revised NEO Personality Inventory (NEO PI-R) and NEO Five Factor Inventory. Professional Manual (Odessa, FL: Psychological Assessment Resources).

Craig, A.D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. Nat. Rev. Neurosci. *3*, 655–666.

de Quervain, D.J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004). The neural basis of altruistic punishment. Science *305*, 1254–1258.

Decety, J., Jackson, P.L., Sommerville, J.A., Chaminade, T., and Meltzoff, A.N. (2004). The neural bases of cooperation and competition: an fMRI investigation. Neuroimage *23*, 744–751.

Delgado, M.R., Stenger, V.A., and Fiez, J.A. (2004). Motivation-dependent responses in the human caudate nucleus. Cereb. Cortex *14*, 1022–1030.

Delgado, M.R., Frank, R.H., and Phelps, E.A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. Nat. Neurosci. *8*, 1611–1618.

Delgado, M.R., Schotter, A., Ozbay, E.Y., and Phelps, E.A. (2008). Understanding overbidding: using the neural circuitry of reward to design economic auctions. Science *321*, 1849–1852.

Eisenberger, N.I., Lieberman, M.D., and Williams, K.D. (2003). Does rejection hurt? An FMRI study of social exclusion. Science *302*, 290–292.

Ellickson, R.C. (2001). The evolution of social norms: a perspective from the legal academy. In Social Norms, M. Hechter and K.D. Opp, eds. (New York: Russell Sage Foundation), pp. 35–75.

Elster, J. (1989). The Cement of Society - A Study of Social Order (Cambridge: Cambridge University Press).

Etkin, A., Egner, T., Peraza, D.M., Kandel, E.R., and Hirsch, J. (2006). Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. Neuron *51*, 871–882.

Fehr, E., and Camerer, C.F. (2007). Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn. Sci. *11*, 419–427.

Fehr, E., and Fischbacher, U. (2003). The nature of human altruism. Nature *425*, 785–791.

Fehr, E., Fischbacher, U., and Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. Hum. Nat. *13*, 1–25.

Fliessbach, K., Weber, B., Trautner, P., Dohmen, T., Sunde, U., Elger, C.E., and Falk, A. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. Science *318*, 1305–1308.

Glimcher, P.W., and Rustichini, A. (2004). Neuroeconomics: the consilience of brain and decision. Science *306*, 447–452.

Gneezy, U. (2005). Deception: The Role of Consequences. Am. Econ. Rev. *95*, 384–394.

Grèzes, J., Frith, C., and Passingham, R.E. (2004). Brain mechanisms for inferring deceit in the actions of others. J. Neurosci. *24*, 5500–5505.

Grèzes, J., Berthoz, S., and Passingham, R.E. (2006). Amygdala activation when one is the target of deceit: did he lie to you or to someone else? Neuroimage *30*, 601–608.

Herwig, U., Baumgartner, T., Kaffenberger, T., Brühl, A., Kottlow, M., Schreiter-Gasser, U., Abler, B., Jäncke, L., and Rufer, M. (2007a). Modulation of anticipatory emotion and perception processing by cognitive control. Neuroimage *37*, 652–662.

Herwig, U., Kaffenberger, T., Baumgartner, T., and Jäncke, L. (2007b). Neural correlates of a 'pessimistic' attitude when anticipating events of unknown emotional valence. Neuroimage *34*, 848–858.

Herwig, U., Brühl, A.B., Kaffenberger, T., Baumgartner, T., Boeker, H., and Jäncke, L. (2009). Neural correlates of 'pessimistic' attitude in depression. Psychol. Med., in press. Published online September 9 2009. 10.1017/S0033291709991073.

Horne, C. (2001). Sociological perspectives on the emergence of norms. In Social Norms, M. Hechter and K.D. Opp, eds. (New York: Russell Sage Foundation), pp. 3–34.

Jäncke, L., Brunner, B., and Esslen, M. (2008). Brain activation during fast driving in a driving simulator: the role of the lateral prefrontal cortex. Neuroreport *19*, 1127–1130.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., and Montague, P.R. (2005). Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. Science *308*, 78–83.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., and Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science *314*, 829–832.

Knoch, D., Nitsche, M.A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., and Fehr, E. (2008). Studying the neurobiology of social interaction with transcranial direct current stimulation—the example of punishing unfairness. Cereb. Cortex *18*, 1987–1990.

Knutson, B., Rick, S., Wimmer, G.E., Prelec, D., and Loewenstein, G. (2007). Neural predictors of purchases. Neuron *53*, 147–156.

Kozel, F.A., Johnson, K.A., Mu, Q., Grenesko, E.L., Laken, S.J., and George, M.S. (2005). Detecting deception using functional magnetic resonance imaging. Biol. Psychiatry *58*, 605–613.

Langleben, D.D., Loughead, J.W., Bilker, W.B., Ruparel, K., Childress, A.R., Busch, S.I., and Gur, R.C. (2005). Telling truth from lie in individual subjects with fast event-related fMRI. Hum. Brain Mapp. *26*, 262–272.

Lee, T.M., Liu, H.L., Chan, C.C., Ng, Y.B., Fox, P.T., and Gao, J.H. (2005). Neural correlates of feigned memory impairment. Neuroimage *28*, 305–313.

Lieberman, M.D. (2007). Social cognitive neuroscience: a review of core processes. Annu. Rev. Psychol. *58*, 259–289.

Liu, X., Powell, D.K., Wang, H., Gold, B.T., Corbly, C.R., and Joseph, J.E. (2007). Functional dissociation in frontal and strFiatal areas for processing of positive and negative reward information. J. Neurosci. *27*, 4587–4597.

Meyer-Lindenberg, A., Buckholtz, J.W., Kolachana, B., Hariri, A., Pezawas, L., Blasi, G., Wabnitz, A., Honea, R., Verchinski, B., Callicott, J.H., et al. (2006). Neural mechanisms of genetic risk for impulsivity and violence in humans. Proc. Natl. Acad. Sci. USA *103*, 6269–6274.

Nuñez, J.M., Casey, B.J., Egner, T., Hare, T., and Hirsch, J. (2005). Intentional false responding shares neural substrates with response conflict and cognitive control. Neuroimage *25*, 267–277.

Phan, K.L., Wager, T., Taylor, S.F., and Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. Neuroimage *16*, 331–348.

Phan, K.L., Magalhaes, A., Ziemlewicz, T.J., Fitzgerald, D.A., Green, C., and Smith, W. (2005). Neural correlates of telling lies: a functional magnetic resonance imaging study at 4 Tesla. Acad. Radiol. *12*, 164–172.

Phillips, M.L., Drevets, W.C., Rauch, S.L., and Lane, R. (2003). Neurobiology of emotion perception I: The neural basis of normal emotion perception. Biol. Psychiatry *54*, 504–514.

Pruessmann, K.P., Weiger, M., Scheidegger, M.B., and Boesiger, P. (1999). SENSE: sensitivity encoding for fast MRI. Magn. Reson. Med. *42*, 952–962.

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., and Kilts, C. (2002). A neural basis for social cooperation. Neuron *35*, 395–405.

Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2004). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. Neuroreport *15*, 2539–2543.

Rilling, J.K., Glenn, A.L., Jairam, M.R., Pagnoni, G., Goldsmith, D.R., Elfenbein, H.A., and Lilienfeld, S.O. (2007). Neural correlates of social cooperation and non-cooperation as a function of psychopathy. Biol. Psychiatry *61*, 1260–1271.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2003). The neural basis of economic decision-making in the Ultimatum Game. Science (New York, N.Y *300*, 1755-1758. Satpute, A.B., and Lieberman, M.D. (2006). Integrating Automatic and Controlled Processes into Neurocognitive Models of Social Cognition. Brain Res. Rev. *1079*, 86–97.

Satpute, A.B., and Lieberman, M.D. (2006). Integrating Automatic And Controlled Processes Into Neurocognitive Models Of Social Cognition. Brain Res. *1079*, 86–97.

Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., and Frith, C.D. (2006). Empathic neural responses are modulated by the perceived fairness of others. Nature *439*, 466–469.

Sip, K.E., Roepstorff, A., McGregor, W., and Frith, C.D. (2008). Detecting deception: the scope and limits. Trends Cogn. Sci. *12*, 48–53.

Spence, S.A., Farrow, T.F., Herford, A.E., Wilkinson, I.D., Zheng, Y., and Woodruff, P.W. (2001). Behavioural and functional anatomical correlates of deception in humans. Neuroreport *12*, 2849–2853.

Spence, S.A., Hunter, M.D., Farrow, T.F., Green, R.D., Leung, D.H., Hughes, C.J., and Ganesan, V. (2004). A cognitive neurobiological account of deception: evidence from functional neuroimaging. Philos. Trans. R. Soc. Lond. B Biol. Sci. *359*, 1755–1762.

Spence, S.A., Kaylor-Hughes, C., Farrow, T.F., and Wilkinson, I.D. (2008). Speaking of secrets and lies: the contribution of ventrolateral prefrontal cortex to vocal deception. Neuroimage *40*, 1411–1418.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., and Fehr, E. (2007). The neural signature of social norm compliance. Neuron *56*, 185–196.

Tabibnia, G., Satpute, A.B., and Lieberman, M.D. (2008). The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychol. Sci. *19*, 339–347.

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations. Econometrica *76*, 1467–1480.

Voss, T. (2001). Game theoretical perspectives on the emergence of social norms. In Social Norms, M. Hechter and K.D. Opp, eds. (New York: Russell Sage Foundation), pp. 105–138.

Zink, C.F., Tong, Y., Chen, Q., Bassett, D.S., Stein, J.L., and Meyer-Lindenberg, A. (2008). Know your place: neural processing of social hierarchy in humans. Neuron *58*, 273–283.