

The Neural Basis of Deception in Strategic Interactions

Kirsten G Volz, Kai Vogeley, Marc Tittgemeyer, D Yves Von_Cramon and Matthias Sutter

Journal Name: Frontiers in Behavioral Neuroscience

ISSN: 1662-5153

Article type: Original Research Article

Received on: 30 Sep 2014

Accepted on: 27 Jan 2015

Provisional PDF published on: 27 Jan 2015

Frontiers website link: www.frontiersin.org

Citation: Volz KG, Vogeley K, Tittgemeyer M, Von_cramon D and Sutter M(2015) The Neural Basis of Deception in Strategic Interactions. *Front. Behav. Neurosci.* 9:27. doi:10.3389/fnbeh.2015.00027

Copyright statement: © 2015 Volz, Vogeley, Tittgemeyer, Von_cramon and Sutter. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](http://creativecommons.org/licenses/by/4.0/). The use, distribution and reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

This Provisional PDF corresponds to the article as it appeared upon acceptance, after rigorous peer-review. Fully formatted PDF and full text (HTML) versions will be made available soon.

The Neural Basis of Deception in Strategic Interactions

Kirsten G. Volz¹, Kai Vogeley^{2,3}, Marc Tittgemeyer⁴, D. Yves von Cramon^{4,5}, and Matthias Sutter^{6,7}

¹*Werner Reichardt Centre for Integrative Neuroscience, Tübingen, Germany*

²*Department of Psychiatry and Psychotherapy, University of Cologne, Germany*

³*Institute for Neuroscience and Medicine – Cognitive Neuroscience (INM3), Research Center Juelich, Germany*

⁴*Max Planck Institute for Metabolism Research, Cologne, Germany*

⁵*Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*

⁶*Department of Public Economics, University of Innsbruck, Innsbruck, Austria*

⁷*Department of Economics, University of Cologne, Cologne, Germany*

Keywords: deception, sophisticated deception, fMRI experiment, temporo-parietal junction, strategic interactions, habenula

Corresponding Author:

Kirsten Volz

Werner Reichardt Centre for Integrative Neuroscience

Otfried-Müller-Straße 25

72076 Tuebingen

Germany

Tel: +49 (0)7071 2989106

Fax: +49 (0)7071 295971

Email: kirsten.volz@cin.uni-tuebingen.de

Abstract

Communication based on informational asymmetries abounds in politics, business, and almost any other form of social interaction. Informational asymmetries may create incentives for the better-informed party to exploit her advantage by misrepresenting information. Using a game-theoretic setting, we investigate the neural basis of deception in human interaction. Unlike in most previous fMRI research on deception, the participants decide themselves whether to lie or not. We find activation within the right temporo-parietal junction (rTPJ), the dorsal anterior cingulate cortex (ACC), the (pre)cuneus (CUN), and the anterior frontal gyrus (aFG) when contrasting lying with truth telling. Notably, our design also allows for an investigation of the neural foundations of sophisticated deception through telling the truth—when the sender does not expect the receiver to believe her (true) message. Sophisticated deception triggers activation within the same network as plain lies, i.e., we find activity within the rTPJ, the CUN, and aFG. We take this result to show that brain activation can reveal the sender's veridical intention to deceive others, irrespective of whether in fact the sender utters the factual truth or not.

Communication based on informational asymmetries abounds in politics, business, and almost any other form of social interaction. Such situations may provide an incentive for either party to exploit the informational asymmetries to their own advantage. This may then imply the use of deception. Although there is some debate about a coherent and generally accepted definition, typically experimental (neuroscientific) investigations are based on a conceptual definition of deception as a deliberate act that is “intended to foster in another person a belief or understanding which the deceiver considers false [...]”. Specifically, the deceiver transmits a false message (while hiding the true information) [...]” (Zuckerman, DePaulo, & Rosenthal, 1981, p. 3). Consider, for example, customers in a restaurant who ask the waiter if the lobster is fresh. The waiter may care only about the customers’ well-being, and answer truthfully. Alternatively, she may be motivated by the restaurant’s need to get rid of the less fresh lobsters and answer untruthfully. Informational asymmetries often provide an incentive for the better-informed party to exploit her informational advantage by holding back information from another party, thus involving some sort of lying or misrepresentation of information.

Yet, wrongly informing the interaction partner about the true nature of a situation is only one form of deception and excludes other important deceptive acts, such as sophisticated deception (Sutter, 2009). By taking into account the sender’s thoughts about the receiver’s belief, sending a true message can also be classified as a form of deception. Particularly, the sender may tell the receiver about the true state of the world, hoping she will think the sender is lying and will therefore not act according to the information provided. For instance, think of opposing parties in war. Here, a sophisticated lie would be to tell the enemy exactly what you are going to do, hoping the opponent will think you are lying and will therefore not act according to the information you provide. In contrast, a plain lie would mean sending the wrong information, such as pretending to invade the other’s territory at a different location from where the attack is actually carried out. Accordingly, sophisticated deception and simple

deception can be delineated along the dimensions “truth of the proposition” (true vs. false) and “the sender’s belief about the receiver’s expectation” (to be deceived vs. not to be deceived), whereas the intention of the sender is in both cases to deceive the receiver. In contrast, sophisticated deception can be delineated from plainly telling the truth along the dimensions “intention of the sender” (to deceive vs. not to deceive) and “the sender’s belief about the receiver’s expectation” (to be deceived vs. not to be deceived). Together, sophisticated deception can be thought of as some sort of a hybrid, it conveys literally the truth, but is intended (and expected) to be perceived as a lie.

In this paper, we analyze the neural foundations of simple as well as sophisticated deception in strategic interactions. Particularly, we ask whether brain activation patterns can reveal the sender’s true intention and can disentangle the two forms of deception, namely simple and sophisticated deception. By using functional Magnetic Resonance Imaging (fMRI), we can derive qualitative and quantitative predictions for brain activation patterns that can help to contrast different candidate strategies that may not be evident from behavioral data alone.

As outlined above and put forward repeatedly for deception in interactive contexts (cp. meta-analysis by Lisofsky et al., 2014), the intention to deceive requires the sender to anticipate the receiver’s mental state and thus think about her beliefs and expectations. Building on the notion that telling the truth is some sort of baseline (Cui et al., 2014), we propose that the intention to deceive the interaction partner, regardless of how it is expressed eventually, requires additional socio-cognitive processes than does telling the truth. This should also be reflected by longer reaction times for both sorts of deceptive behavior when compared to truth telling as well as be reflected on the phenomenological level (i.e., senders’ reports). Therefore, we expect increased neural activation when comparing simple and sophisticated deception to plainly telling the truth specifically within regions that have been associated with theory of mind (ToM) processes, such as the right temporo-parietal junction

(TJP), including the posterior superior temporal gyrus/angular gyrus (Amodio & Frith, 2006; Decety & Lamm, 2007; Frith & Frith, 1999; Vogeley et al. 2001; Wolf, Dziobek, & Heekeren, 2010) and with social cognition, such as the temporal pole (TP) (Frith, 2007; Moriguchi et al., 2006; for a review see Olson et al., 2007). The hypothesized activation pattern reflecting the intention to deceive (TPJ, TP,) shall also be observed for sophisticated deception when compared to plain truth trials. Therefore, we could distinguish the two forms of sending objectively true messages and unfold the sender's true (deceptive) intention. Finding activation within areas reflecting socio-cognitive processes specifically for deceptive behavior (irrespective of how it unfolds) as compared to truth telling would be novel and taken to indicate the specific requirement of such processes for deception in strategic interactions. In other words, if the outcome of the interaction depends on both, the sender and the receiver, deceptive behavior – undertaken to get a (monetary) advantage – requires other processes than solely saying the truth and therewith accepting the outcome of the interaction without any attempts to influence it.

For plain lies (as compared to plainly telling the truth) we expect (in addition to TPJ and TP) activation within the anterior cingulate cortex (ACC). A recent quantitative meta-analysis on deceptive behavior in social interactive paradigms (Lisofsky et al., 2014) suggested this activation “to indicate greater conflict processing during deception in social situations in which people are especially supposed to behave honestly” (Liskofsky et al., 2014, p. 119). This ACC activation for plain lies is expected to vary depending on the intensity of conflict, which we define as the product of the differences of the sender's and the receiver's monetary payoffs.

Taking into account the sender's true intention, allows us (for the first time) to specifically investigate the neural correlates of genuine truth trials. In none of the previous imaging studies on deception did the authors report any specific activation pattern for telling the truth. If this was due to truth trials being a heterogeneous category (for instance, including

truth trials with the intention to deceive), we shall find a specific activation pattern for telling the truth in this study.

Studying deception in strategic interaction requires participants be given a choice of whether to deceive another person, because only when they have a choice can we find out the circumstances under which subjects will resort to deception (Abe et al., 2007; Greene & Paxton, 2009; Sip et al., 2010; for a review see Sip, Roepstorff, McGregor, & Frith, 2008). For this reason, paying subjects according to their choices – as is standard in experimental economics (Smith, 1976) – is important. Accordingly, in the present study participants played a simple sender-receiver game (Crawford & Sobel, 1982; Gneezy, 2005). In this two-person game, the sender (e.g., the waiter in the introductory example) is informed about two possible states of the world (the lobster is fresh or not) that yield particular payoffs for the sender and the receiver (the customer). The sender can send a message to the receiver that is either true or false with respect to which state of the world is more profitable for the receiver. Based on this message, the receiver makes a decision (whether to order the lobster or not), thus determining the payoffs for the sender and the receiver. That is, the monetary payoff for the sender highly depends on whether she is successful in making the receiver believe her. We assume the receiver cannot figure out whether the message is true (e.g., the customer cannot retaliate if he finds out the lobster was bad). This is different from a recent paper on the neural circuitry of a broken promise in which the person sending a promise was also the person making the decision about whether to keep the promise (Baumgartner et al., 2009). In our context, sending a message is the only action the sender can take and thus the only way in which she might influence the receiver. Taken together, our paradigm addresses widespread concerns around ecological validity of experiments on deception in that it is truly interactive, participants have a real opportunity to deceive another person who is not a confederate, and participants' payoffs (in the role of the sender) depend fully on the decision of the receiver.

Moreover, due to the specificity that the receiver cannot find out whether the sender had sent a wrong message or not allows us to investigate deceptive behavior in strategic interactions that is unaffected by learning and adaptation effects. It is for the latter reason that we give no feedback to receivers about the actual options from which the sender could choose from.

Materials and Methods

Participants

Thirty-four (17 women, mean age 24.3 years, SD 2.6, range 21-32 years) right-handed¹, healthy volunteers (without any neurological or psychiatric history) participated in the fMRI experiment for a payment of 12 Euro per hour. Additionally to this show-up fee, participants could earn up to 30 Euros. That is, at the end of the experiment, one trial was randomly drawn and paid out according to the receiver's choice in this specific trial. All participants had normal or corrected-to-normal vision, spoke German as their native language, and none had irremovable metal implants in their bodies. The experimental procedure and data collection followed the ethical guidelines of the "Declaration of Helsinki" (revised version, 2012) and were approved by the local ethical committee of the University of Cologne. Data were handled anonymously. We had to exclude four participants (1 male and 3 female) from the analysis because of too few lying or sophisticated deception trials, respectively, and one participant because of zero truth trials.

Stimuli and Experimental Paradigm

In the sender-receiver game, there are two players of which only the sender (the person being scanned) is informed about the monetary consequences for herself and the receiver for two different options, one being associated with Blue color and the other with Red color. Let Blue

¹ In recent years, a vast number of imaging studies have shown that there are marked differences in the neural localization of cognitive (and especially language) functions in the brains of left-handed individuals when compared with right-handers. To exclude a putative effect of lateralization correlated with handedness we had chosen to only include right-handers in our imaging study.

(S^b, R^b) represent the payoff to the sender and the receiver, respectively, from choosing Blue, and Red (S^r, R^r) from choosing Red (cp. Figure 1). After being informed about these pairs of payoffs, the sender sends a message to the receiver, saying either “Blue is more profitable for you” or “Red is more profitable for you.” After sending a message, the sender has to indicate on a new screen which state she expects the receiver to pick. Then the next trial started. All in all, 90 games were played that differed with respect to the relative gains and losses for the two players (see below).

We call a choice a *sophisticated deception* when a sender sent the true message expecting the receiver *not* to follow it. We denote as a *true message* a case in which a sender sent the true message and expects the receiver to follow it by picking the state the message indicated as more profitable. We classify as *simple deception* cases in which the sender sent the false message. After receiving the message from the sender, the receiver chooses Blue or Red, and the respective payoffs are recorded (cp. Table 1).

While the sender underwent the anatomical scanning session (to obtain the individual anatomical structures onto which the metabolic activity map was projected), the receiver played the game in another room, which was located across campus, and it was ensured that sender and receiver never met each other. This was done to exclude any effects that might arise as a consequence of attractiveness, sympathy, gender, or the like. After the receiver finished her part (which was approximately at the time the scanning session of the sender was finished), one trial was randomly picked by the experimenter and the corresponding payoff (additional to the show-up fee of 12 Euros/h) was paid out to the sender and the receiver according to the receiver’s choice. The mean additional payout for senders was €8.53 (SD=5.54), ranging from 5 to 20 Euros; the mean additional payout for receivers was €8.26 (SD=6.15), ranging from 0 to 25, not being significantly different ($t(58)=-.176, p=.861$). The full set of instructions is provided in the Appendix and both, sender and receiver, knew about the entire procedure before starting the experimental session (see A.1 and A.2).

In each of the 90 games, the sender was asked to send one of the above messages to the receiver. One of these messages was always true and the other was false. Knowing only the message she received and not the potential payoffs in each state, the receiver had to pick either Blue or Red, which then determined the payoffs for the sender and the receiver. Since the receiver was only informed about her actual payoff in the chosen state—and not about the sender’s actual payoff or the possible payoffs in the un-chosen state—the receiver could not judge whether the sender had told the truth or not. Yet, the receiver was informed that the maximum profits for her and the sender was 30 Euros. It was important that the receiver did not know about the potential payoffs in each state (but only the payoff of the actually chosen option in the current trial), otherwise she would have adjusted her behavior, thus confounding objectivity and comparability (within and across participants) as well as affecting the sender’s strategic behavior. Likewise, to exclude learning and order effects on the side of the sender’s behavior, the sender did not learn about the decisions of the receiver.

We varied the incentives for deception along three different categories for the 90 games, indicating the possible tension between the sender’s and receiver’s payoffs (i.e., stimulus-dependent categorization independent of participants’ choice). In the category “*conflict*” (n=45), the more profitable state for the sender was always less profitable for the receiver. We also varied the relative gains and losses of the sender and the receiver between the two states of a game. In category “*sender indifferent*” (n=27), the sender earned the same amount of money in both states, but the receiver payoff differed across states, and it could be higher or lower than the sender’s payoff. Category “*aligned interests*” (n=18) included only pairs of states in which one state yielded higher profits both for the sender and the receiver, although the increase in payoffs from the worse to the better state could differ for sender and receiver. The order of presentation of games was randomized. The full set of games is provided in the Appendix (see A.3).

All trials lasted for 16 s (i.e., 8 scans at $TR = 2$ s): the game with its monetary payoffs was presented for a maximum of 8 s, during which time participants could respond, followed by a short fixation (2 s) and then the question about the sender's expectation (4 s). Subsequently, the announcement that the next trial was about to start was presented for 2 s. To help to characterize the shape of the hemodynamic response function, the timing of the presentation of the stimulus was varied. Accordingly, using a jittering-method more points of the hemodynamic response function can be sampled than if a fixed inter-stimulus-interval was used. Particularly, we randomly varied the onset of each stimulus presentation relative to the beginning of the first of the eight scans (0, 400, 800, 1200, 1600 ms) to enhance the temporal resolution of the signal captured (Birn, Cox, & Bandettini, 2002; Miezin et al., 2000).

MR Scanning Procedure

Image Acquisition: Imaging was performed on a 3T scanner (Siemens TRIO, Erlangen, Germany) equipped with a standard birdcage head coil. Participants lay supine in the scanner with their hands placed on a right and left response button box. The index fingers were placed on two appropriate response buttons and participants were trained about the response contingencies. Form-fitting cushions were used to prevent participants from head movement and they were provided with earplugs to attenuate the scanner noise. The experiment was presented via a mirror that was mounted to the headcoil and individually adjusted.

One of the areas, in which we expected activation, is the temporal pole. This area is subject to severe distortion and signal loss in fMRI due to susceptibility artifacts that result from the area's specific location, i.e., near air-filled sinuses (Ojemann et al., 1997). Therefore, we used a spin-echo (SE) sequence which has been shown to be less sensitive to susceptibility-related signal dropouts as in contrast to gradient-echo (GE) sequences (Schmidt, Boesiger, & Ishai, 2005; Norris et al., 2002). Yet, the drawback of using SE-based instead of GE-based fMRI is a lower statistical power of the SE sequences.

During functional imaging, 17 axial slices (4 mm thickness, 25% spacing, field of view [FOV] 21 cm, data matrix of 64 x 64 voxels, and in-plane resolution of 3.3 mm x 3.3 mm) covering the whole brain were collected using a single-shot spin-echo echo-planar imaging (SE-EPI) sequence (TR 2 s, echo time [TE] 80 ms, flip angle 90°) sensitive to blood oxygen level-dependent (BOLD) contrast. One functional run with 728 timepoints was run with each time point sampling over the 17 slices. After the functional imaging, high-resolution 3D T1-weighted whole brain MDEFT sequences (128 sagittal slices, 1 mm thickness) were recorded.

Image processing and analysis: The functional imaging data were processed and analyzed using the software package LIPSIA (Leipzig Image Processing and Statistical Inference Algorithms) version 2.2 (Lohmann et al., 2001). To correct for temporal offsets between the slices acquired in one scan, a cubic-spline interpolation was used. Thereafter the data were motion-corrected with the 50th time-step as a reference and 6 degrees of freedom (3 translational, 3 rotational). A temporal high-pass filter with a cutoff frequency of 1/120 Hz was used to remove low-frequency signal changes and baseline drifts and a spatial Gaussian filter with 6 mm full-width half-maximum (FWHM) was applied. A rigid linear registration with six degrees of freedom (three rotational, three translational) was performed to align the functional data slices with a 3D stereotactic coordinate reference system. The rotational and translational parameters were acquired on the basis of the MDEFT slices to achieve an optimal match between these slices and the 3D reference data dataset. The MDEFT volume data was standardized to the MNI atlas. The rotational and translational parameters were subsequently transformed by linear scaling to the same standard size. The resulting parameters were then used to transform the functional slices employing a trilinear interpolation, so that the resulting functional slices were aligned with the stereotactic coordinate system. Resulting data had a spatial resolution of 3 x 3 x 3 mm (27mm³).

The statistical evaluation was based on a least-squares estimation using the general linear model (GLM) for serially auto-correlated observations (Friston et al., 1995; Worsley & Friston, 1995). The design matrix was generated with a delta function, convolved with the hemodynamic response function (gamma function) (Glover, 1999). We used two different design matrices to answer the different research questions. One design matrix comprised the following events: truth trials, simple deception trials, and sophisticated deception trials (cp. Table 1). The trials were classified based on participants' behavior, i.e., their choice which message to send to the receiver and their response to the question "Which state do you expect the receiver to choose?". Events were modeled time-locked to the beginning of a game. The duration was modeled individually with the time it took participants to respond to the game (RT) (Grinband et al., 2008) and with amplitude of one. In another design matrix that was used to model and investigate the effects of conflict (defined as the tension between the sender's and receiver's payoffs), we had five regressors, particularly, truth trials, simple deception trials, and sophisticated deception trials with their duration being modeled individually by RT and amplitude of one plus two regressors for simple deception trials and sophisticated deception trials that were modeled with their individual RT and an amplitude that reflected the tension between the sender's and receiver's payoffs. The tension to deceive was calculated as the product of the differences of the sender's and the receiver's payoff for the pairs of states, i.e., $(S^b - S^r) * (R^r - R^b)$. (cp. description of stimulus material and Figure 1). For instance, let $S^b=15$, $S^r=5$, $R^r=15$, and $R^b=5$, then the value representing the tension between the player's payoffs is $(15-5)*(15-5) = 100$. In contrast, for a matrix with the payoffs $S^b=1$, $S^r=0$, $R^r=5$, and $R^b=0$, the conflict value is low $((1-0)*(5-0)=5)$. This value represents the product of the difference of the profit of the sender and the corresponding inverted difference of the receiver. This means that if the differences have opposite signs, then the sender and the receiver have conflicting interests. In case the differences have the same sign, both the sender and the receiver gain higher profits in the same state. If the sender is indifferent between the

two states, the parameter value is zero. Hence, this conflict parameter reflects a measure of the tension to deceive.

For each participant, contrast images were generated on the basis of beta-value estimates of the raw-score differences between specified conditions. Subsequently, these single subject contrasts were entered into a second-level analysis on the basis of Bayesian statistics (Lazar, 2008; Neumann & Lohmann, 2003). In the approach by Neumann and Lohmann (2003), posterior probability maps and maps of the effect size are calculated on the basis of the resulting least-square estimates of parameters for the GLM. That is, the parameter estimates on the second level of analysis are viewed within a Bayesian framework as evidence for the presence or absence of the effect of interest in a group of participants. The output of the Bayesian second-level analysis is a probability map showing the probability for the contrast to be larger than zero. This Bayesian technique allows us to directly estimate the probability of a specific difference in the group means given the parameter estimates of the GLM for the individual participants. This is more informative than a classical rejection of a null hypothesis. This approach has the further advantage, when compared with conventional analyses based on t statistics, of being less sensitive to outliers than traditional t statistics, as the influence of individual participants on a group statistic is weighted by the within-subject variability. In support of this, Thirion et al. (2007) suggested that, from the point of view of reliability, optimal statistical thresholds for activation maps are lower than classical thresholds corrected for multiple comparisons. Furthermore, since probabilities of the contrasts are calculated, but no significance tests are performed, corrections for multiple comparisons or calculations of effect sizes are not necessary. For visualization, a threshold of 99.4% was applied to the probability maps.

Results

Behavioral Results

As expected, the frequency of sending the false message strongly depends on a game's category, i.e., on the distribution of payoffs (for a description of the stimulus-dependent categorization please see Stimuli and Experimental Paradigm): it is fairly low in the "*aligned interest*" category (25%, SD 22.5) and in "*sender indifferent*" (24.7%, SD 23.2), but comparatively high in "*conflict*" (60.8%, SD 21.5) ($F(2,28)=34.97, p=.0001$). Lying in the "*conflict*" category is significantly more frequent than in either "*aligned interest*" or "*sender indifferent*", whereas we find no significant difference between the latter two categories. Furthermore, the possible gains for the sender if the receiver picks the state that is better for the sender, and the potential losses for the receiver if she picks the state that is worse for her, have a significant impact on the likelihood of sending the false message. Senders lie more often when the potential gains from lying are high (10€ or 5€; 55.8%, SD 19.2) than when they are low (1€ or 0€; 34.2%, SD 17.2; $t(29)=6.1, p=.0001$). Senders lie less often when the possible losses for the receiver are high (10€, 15€ or 20 €; 37.4%, SD 20.3) than when they are low (1€ or 5 €; 47.2%, SD 13.3; $t(29)=-4.04, p=.0001$). These results clearly indicate that monetary incentives affect the frequency of sending the false message.

The relative frequency of sophisticated deception (as a fraction of the total number of cases in which the sender sent the objectively true message) depends on a game's category in the same way the frequency of simple deception does. In the category "*conflict*", we observe sophisticated deception in 59.3% (SD 31.5) of cases with true messages, whereas we observe it significantly less often in "*sender indifferent*" (40.9%, SD 28.6) and "*aligned interest*" (31.7%, SD 19.1; $F(2,26)=14.98, p=.0001$). This finding indicates sophisticated deception through telling the truth is most likely when the sender can profit most from it. Adding the cases of sophisticated deception to the cases of simple deception, the overall frequency of deception reaches 65.9% (SD 18.9) across all categories, whereas it is only 42.8% (SD 15.3) when taking into account only plain lies and ignoring deceptive behavior through truth telling.

Our assumption that truth telling may be less demanding than deceiving the interaction partner was confirmed for both sorts of lying: Telling a plain lie ($M=2618$ ms, $SD=202$) or engaging in sophisticated deception ($M=2611$ ms, $SD=193$) – while not significantly different from each other – takes significantly longer than telling the truth ($M = 2453$ ms, $SD = 211$) ($F(2,30)=3.46$, $p=.044$). This response pattern is crucially affected by the actual payoffs: A 3 (category) x 3 (response (truth, SD, plain lies)) repeated measures ANOVA reveals a significant main effect of category ($F(2,30)=6.44$, $p=.005$), with “conflict” trials showing the longest RTs ($M=2720$ ms, $SD=211$) followed by “sender indifferent” trials ($M=2565$ ms, $SD=203$) and “aligned interest” trials ($M=2397$ ms, $SD=200$) (cp. Table 2). We take these results to support the notion that deceptive behavior, irrespective of how it is expressed, demands additional cognitive processes so as to suppress a pre-potent truthful answer. This is also supported by our post-session questionnaire data: senders report that it took them significantly longer to respond when stakes were high and that they had to deliberate harder when preparing to deceive the receiver.

Additional results from the post-session questionnaire data reveal insights regarding strategy and heterogeneity. Concerning the former, 86.6% of the senders report having developed a strategy how to interact with the receiver and of those more than half (59.9%) report that their strategy depended on the difference in payoffs between sender and receiver as well as on the absolute amounts. The remaining senders indicate to have taken into account the frequency and succession of previous blue- and red-responses so as to determine how to respond. We take these findings to indicate that senders engaged, indeed, in our social interactive paradigm and cared about the actual payoffs. Concerning the issue of heterogeneity, the data display a heterogeneous sample. Being asked on how many of the trials they produced a deceptive response, senders on average say that they did so in 43.8% ($SD=23$) of the cases, the range being 5% to 90%. A closer look reveals that 36.6% of the senders have had a bad conscience when producing a deceptive response (with the feeling

even persisting for a couple of trials) and feel that they had lied in effect. These senders indicate to have lied in only a third of the trials ($M=33.4\%$, $SD=21.6$). In contrast, the other senders (63.3%) report not having had a feeling of actually lying, and thus indicate having lied in approximately half of the trials ($M=49.7\%$, $SD=22.1$, $t(28)=-1.95$, $p=.06$ (2-tailed)).

Imaging Results

Neural correlates of the intention to deceive in strategic interactions (simple and sophisticated deception > truth)

To study the neural correlates of the intention to deceive, we contrast the hemodynamic activation of simple deception trials *and* sophisticated deception trials with truth trials and find activation within the right TPJ, superior temporal gyrus, precuneus extending into the retrosplenial cortex, cuneus bilaterally, and within the right superior frontal gyrus (BA 10) (see Table 3 and Figure 2, upper panel).

Neural correlates of lying in strategic interactions (simple deception > truth)

To study the neural correlates of simple deception, i.e., sending a false message with the intention to deceive, we contrast the hemodynamic activation of simple deception trials with truth trials and find activation within the right TPJ, the dorsal ACC, the precuneus extending into the retrosplenial cortex, within the cuneus, the right anterior frontal gyrus, and a comparatively small activation focus within the anterior medial prefrontal cortex (amPFC) (see Table 4 and Figure 3, upper panel).

Neural correlates of sophisticated deception (sophisticated deception > truth)

To study the neural correlates of sophisticated deception specifically, we built a contrast of sophisticated deception trials and truth trials. We find activation within the right TPJ, the

precuneus, the left cuneus, the right anterior frontal gyrus (BA 10), and the superior temporal gyrus (see Table 5 and Figure 3, lower panel).

Importantly, this finding suggests sophisticated deception is not a variant of plainly telling the truth – in which case no activation differences in this contrast should have occurred – but a version of telling a lie, since a very similar activation pattern occurred as in the contrast “simple deception versus truth” (cp. upper panel in Figure 3).

Neural correlates delineating the two forms of deception (sophisticated deception > simple deception)

To test for the differences between the two forms of deception, we contrasted sophisticated deception trials with simple deception trials. We find activation bilaterally within the TPJ, the right middle temporal gyrus, the left superior temporal gyrus, the left frontal operculum, and within the mid-cingulate gyrus (see Table 6 and Figure 4, upper panel).

Neural correlates of genuine truth trials

Taking into account the sender’s true intention, we are able to extract genuine truth trials, i.e., trials where the sender sent the true message with the expectation that the receiver believes her (true) message. These trials are contrasted with both simple deception as well as sophisticated deception trials. We find activation within the habenular complex bilaterally, the right frontal operculum, the left pregenual ACC, and the right middle frontal gyrus (see Table 7 and Figure 2, lower panel).

Parametric analysis modeling the incentive to deceive for simple deception trials

To test whether the activation that revealed for simple deception varies with the monetary incentive, we calculate a parametric analysis. Responses are modeled by a value that reflects the tension between the sender’s and the receiver’s payoffs. It is calculated as the product of

the differences of the sender's and the receiver's payoff for the pairs of states, i.e., $(S^b - S^r) \cdot (R^r - R^b)$ (cp. MR Scanning Procedure). The posterior probability maps of this parametric analysis reveals the anterior median prefrontal cortex (amPFC), the dorsal anterior cingulate cortex (ACC), and the anterior frontal gyrus (BA 10) to be more engaged the higher the conflict and thus the tension in payoffs between sender and receiver (see Table 8 and Figure 4, lower panel).

Discussion

Many real life situations are characterized by informational asymmetries among interacting parties. Obviously, such situations may provide an incentive for either party to exploit the informational asymmetries to their own advantage. This may then imply the use of deception. In this fMRI study we analyze the neural foundations of deception in strategic interactions. Notably, in our paradigm, interaction partners were free whether or not to lie. Besides plain lying, we study a broader concept of deception by looking at what has been called sophisticated deception (Sutter, 2009). Here, telling the truth is counted as an act of deception when the sender expects the receiver not to follow the sender's (true) message. Moreover, by taking into account the sender's true intention, we can also determine the neural correlates of genuine truth trials. All in all, we take our results to show that brain activation patterns can reveal the sender's true intention (to deceive), for instance when sending an objectively true message.

Intention to deceive

Particularly, our results reveal the right temporo-parietal junction (rTPJ), the (pre)cuneus (CUN), retrosplenial cortex, and anterior frontal gyrus (aFG) to be specifically involved for the intention to deceive, irrespective of whether this is done by sending a false or a true message. The finding of activation within the rTPJ is in line with our hypothesis. Based on

previous findings and recent meta-analytic findings on deceptive behavior, we suggest this activation to reflect socio-cognitive processes during deception. Specifically, deceptive behavior crucially depends on the ability to anticipate the receiver's mental state. The rTPJ, including posterior superior temporal and angular gyrus, have repeatedly been shown to be specifically involved when people have to integrate socially relevant information and to infer the mental states of others (Bahnemann et al., 2010; Saxe & Kanwisher, 2003; Decety, & Grèzes, 2006; Decety, & Lamm, 2007; Saxe, 2006). Thus, the finding of rTPJ activation for deceptive behavior, realized either by telling a lie or telling the truth, is consistent with our hypothesis on the intentional aspects of deception in a social setting, in which the intentional states of others are integrated into one's own reasoning (Grèze, Frith, & Passingham, 2004; Perner et al., 2006; Saxe, in press; Saxe & Kanwisher, 2003; Walter et al., 2004).

Activation within the cuneus, precuneus and aFG were not expected specifically but cuneus activation may reflect increased requirements as to early visual processing (Vanni et al., 2001), e.g., when thoroughly inspecting the payoff matrix, that is then sent to several parietal areas (Fattori, Pitzalis, & Galletti, 2009); precuneus activation may reflect increased episodic memory retrieval processes (Cavanna & Trimble, 2006), for instance, retrieving past payoff matrices and one's choices in the sender-receiver game, as well as automatic social monitoring processes when observing interacting people (Iacoboni et al., 2004; Leube et al., 2012; Vrticka et al., 2013). And activation within the aFG may reflect the integration of the outcomes of two separate cognitive operations in the pursuit of a higher goal (Ramnani & Owen, 2004).

Deception through telling the truth (sophisticated deception)

Notably, finding this activation pattern both for simple as well as sophisticated deception trials, reveals that sophisticated deception, although superficially appearing as truth trials, cannot be considered a variant of plainly telling the truth – in which case no activation

differences between sophisticated deception and truth trials should have occurred. Rather, the intention to deceive seems to share a lot with deceptive behavior in terms of cognitive processes. Sophisticated deception, as defined in the context of our sender-receiver game, is a form of deception that crucially has to take into account the receiver's reasoning. The sender has to form expectations about the receiver's beliefs and has to adjust her own actions accordingly. Hence, rTPJ activation becomes characteristic for sophisticated deception. Based on this finding, we suggest that brain activation can reveal the sender's veridical intention to deceive in the absence of overt lying. Accordingly, it seems warranted not to confine deception simply to telling a lie.

Interestingly, sophisticated deception seems also to stand out from simple deception. That is, trying to deceive the interaction partner by telling the truth requires greater processing demands than simply telling a lie. Particularly, given activation within the TPJ, ISTG, and MTG, we take this result to indicate greater demands when reading or inferring the partner's thoughts and beliefs so as to correctly predict the receiver's actions. That is, sophisticated deception differs from plainly telling a lie by heightened demands for ToM processes. Instead of construing additional activation (for instance within the frontal gyrus), our result may be understood as representing increasingly more complex processing of the social situation in strategic interaction (Bahnemann et al., 2010).

A further indication that simple and sophisticated deception are two different forms of deceptive behavior come from the parametric analysis. Only for simple deception trials part of the respective network was modulated by the distribution of monetary payoffs between sender and receiver. That is, activation within the dorsal ACC, amPFC, and aFG correlated positively the higher the conflict between sender's and receiver's payoffs. Activation within the dACC has consistently been related to conflict detection and monitoring processes (Carter & van Veen, 2007), although "conflict monitoring may be just one facet of the broader role of ACC in performance monitoring and the optimization of behavior" (Yeung, 2013, p.1, in press).

Carter and van Veen (2007) suggested the ACC's specific role is "to detect conflict between simultaneous active, competing representations and to engage the dorsolateral prefrontal cortex (DLPFC) to resolve such conflict" (p. 367). The greater involvement of this area for high conflict trials when sending false messages may indicate greater tension in situations where people resort to lying despite knowing of the normative appeal to tell the truth.

Genuine truth trials

By taking into account the sender's veridical intention, we could determine the neural correlates of genuine truth trials in the present study. Hitherto, imaging studies on deceptive behavior did not report any significant activation for telling the truth, which could be due to truth trials being a heterogeneous category in those studies, potentially also including sophisticated deception trials. We found significant activation within the habenular complex bilaterally and the left frontal operculum and MTG. Based on animal research, the habenular complex has been suggested to be specifically involved in the control of the human reward system. For instance, the electrical stimulation of the habenular nuclei resulted in an inhibition of up to 90% of the dopamine neurons in the ventral tegmental area and substantia nigra in rats (Christoph, Leonzio, & Wilcox, 1986). In contrast, lesions to the habenular complex resulted in an "increased dopamine turnover in the nucleus accumbens, striatum, and prefrontal cortex, reflecting an activation of the dopaminergic system (Lisoprawski et al., 1980; Nishikawa, Fage, & Scatton, 1986)" (Ullsperger & von Cramon, 2003, p. 4308). Based on these as well as anatomical data, it has been suggested that the habenular complex serves as a "critical modulatory relay between the limbic forebrain structures and the midbrain" (Ullsperger & von Cramon, 2003, p. 4309). Accordingly, habenular activation for telling the truth in strategic interactions in the present study may reduce the probability of phasic dopamine release in the reward system, and thus may reinforce truth telling through weakening the incentive of the monetary profits.

In sum, our study provides a new paradigm for studying the neural basis of deception in human interaction. Contrary to previous studies with instructed deception in non-interactive contexts, we have created a social interactive context based on game-theoretic modeling. Importantly, we are the first to investigate the neural foundations of an intention to deceive in the absence of overt lying. Such sophisticated deception through telling the truth is an intriguing alternative to telling a plain lie, and it can be strategically used, as in the Austrian writer Franz Grillparzer's comedy "Woe to him who is lying" in which the young kitchen boy Leon frees his bishop's captured nephew by telling the guards he is going to free their hostage, and they let him proceed because they don't believe him.

Acknowledgment

We thank Caroline Szymanski, Thomas Dratsch, Philipp Euskirchen, Volker Neuschmelting and Laura Mega for programming the experimental paradigm and for help in data analysis, and Hilke Plassmann as well as three anonymous reviewers for their helpful and constructive comments on an earlier draft of this manuscript.

References

- Abe, N., Suzuki, M., Mori, E., Itoh, M., & Fujii, T. (2007). Deceiving others: Distinct neural responses of the prefrontal cortex and amygdala in simple fabrication and deception with social interactions. *Journal of Cognitive Neuroscience*, 19, 287-295.
- Amodio, D.M., & Frith, C.D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Review Neuroscience*, 7, 268-277.
- Bahnemann, M., Dziobek, I., Prehn, K., Wolf, I., & Heekeren, H. R. (2010). Sociotopy in the temporoparietal cortex: Common versus distinct processes. *Social Cognitive, and Affective Neuroscience*, 5, 48-58.
- Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., & Fehr, E. (2009). The neural circuitry of a broken promise. *Neuron*, 64, 756-770.
- Birn, R.M., Cox, R.W., & Bandettini, P.A. (2002). Detection versus estimation in event-related fMRI: Choosing the optimal stimulus timing. *NeuroImage*, 15, 252-264.
- Carter, C. S., & van Veen, V. (2007). Anterior cingulate cortex and conflict detection: An update of theory and data. *Cognitive, Affective, and Behavioral Neuroscience*, 7, 367-379.
- Cavanna, A. E., Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain*, 129, 564-583.
- Christoph, G. R., Leonzio, R. J., & Wilcox, K. S. (1986). Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. *Journal of Neuroscience*, 6, 613-619.
- Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50, 1431-1451.
- Decety, J., & Grèzes, J. (2006). The power of simulation: Imagining one's own and other's behavior. *Brain Research*, 1079, 4-14.

- Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *Neuroscientist*, 13, 580-593.
- Fattori, P., Pitzalis, S., & Galletti, C. (2009). The cortical visual area V6 in macaque and human brains. *Journal of Physiology – Paris*, 103, 88-97.
- Friston K. J., Holmes A. P., Worsley K. J., Poline J. P., Frith C. D., Frackowiak R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2, 189–210.
- Frith, C. D. (2007). The social brain? *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 362, 671–678.
- Frith, C.D., & Frith, U. (1999). Interacting minds – a biological basis. *Science*, 286, 1692-1695.
- Glover G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage* 9, 416–429.
- Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95, 384-394.
- Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, 106, 12506-12511.
- Grèze, J., Frith, C.D., & Passingham, R.E. (2004). Inferring false beliefs from the actions of oneself and others: an fMRI study. *NeuroImage*, 21, 744-750.
- Grinband, J., Wager, T. D., Lindquist, M., Ferrera, V. P., & Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *NeuroImage*, 15, 509-520.
- Iacoboni M., Lieberman M. D., Knowlton B. J., Molnar-Szakacs I., Moritz M., Throop C. J., et al. (2004). Watching social interactions produces dorsomedial prefrontal and medial

- parietal BOLD fMRI signal increases compared to a resting baseline. *NeuroImage*, 21, 1167–1173.
- Lazar, N. (2008). The statistical analysis of functional MRI data. New York: Springer Verlag.
- Leube, D., Straube, B., Green, A., Blumel, I., Prinz, S., Schlotterbeck, P., et al. (2012). A possible brain network for representation of cooperative behavior and its implications for the psychopathology of schizophrenia. *Neuropsychobiology*, 66, 24–32.
- Lisofsky, N., Kazzer, P., Heekeren, H. R., & Prehn, K. (2014). Investigating socio-cognitive processes in deception: A quantitative meta-analysis of neuroimaging studies. *Neuropsychologia*, 61, 113-122.
- Lisoprawski, A., Herve, D., Blanc, G., Glowinski, J., & Tassin, J. P. (1980). Selective activation of the mesocortico-frontal dopaminergic neurones induced by lesions of the habenula in the rat. *Brain Research*, 183, 229-234.
- Lohmann, G., Muller, K., Bosch, V., Mentzel, H., Hessler, S., Chen, L., et al (2001). LIPSIA—A new software system for the evaluation of functional magnetic resonance imaging of the human brain. *Computerized Medical Imaging and Graphics*, 25, 449-457.
- Miezin, F.M., Maccotta, L., Ollinger, J.M., Petersen, S.E., & Buckner, R.L. (2000). Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on their relative timing. *NeuroImage*, 11, 735-759.
- Moriguchi, Y., Ohnishi, T., Lane, R. D., Maeda, M., Mori, T., Nemoto, K., et al. (2006). Impaired self-awareness and theory of mind: an fMRI study of mentalizing in alexithymia. *NeuroImage*, 32, 1472–1482.
- Neumann, J., & Lohmann, G. (2003). Bayesian second-level analysis of functional magnetic resonance images. *NeuroImage*, 20, 1346-1355.

- Nishikawa, T., Fage, D., & Scatton, B. (1986). Evidence for and nature of the tonic inhibitory influence of the habenulointerpenduncular pathway upon cerebral dopaminergic transmission in the rat. *Brain Research*, 373, 323-336.
- Norris, D.G., Zysset, S., Mildner, T., & Wiggins, C.J. (2002). An investigation of the value of spin-echo-based fMRI using a Stroop color-word matching task and EPI at 3T. *NeuroImage*, 15, 719-726.
- Ojemann, J. G., Akbudak, E., Snyder, A. Z., McKinstry, R. C., Raichle, M. E., & Conturo, T. E. (1997). Anatomic localization and quantitative analysis of gradient refocused echo-planar fMRI susceptibility artifacts. *NeuroImage*, 6, 156–167.
- Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: a review of findings on social and emotional processing. *Brain*, 130, 1718–1731.
- Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience*, 1, 245-258.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 4, 515-526.
- Ramnani, N., & Owen, A.M. (2004). Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Review Neuroscience*, 5, 184-194.
- Saxe, R. (in press). Theory of Mind (Neural Basis). *Encyclopedia of Consciousness*.
- Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, 16, 235-239.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in “theory of mind”. *NeuroImage*, 19, 1835–1842.
- Schmidt, C. F., Boesiger, P., & Ishai, A. (2005). Comparison of fMRI activation as measured with gradient- and spin-echo EPI during visual perception. *NeuroImage*, 26, 852-859.

- Sip, K. E., Roepstorff, A., McGregor, W., & Frith, C.D. (2008). Detecting deception. The scope and limits. *Trends in Cognitive Science*, 12, 48-53.
- Sip, K. E., Lynge, M., Wallentin, M., McGregor, W. B., Frith, C. D., et al. (2010). The production and detection of deception in an interactive game. *Neuropsychologia*, 48, 3619-3626.
- Smith, V. L. (1976), Experimental economics: Induced value theory. *American Economic Review*, 66, 274–279.
- Sutter, M. (2009). Deception through telling the truth?! Experimental evidence from individuals and teams. *The Economic Journal*, 119, 47-60.
- Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S. et al., (2007). Analysis of fMRI data sampled from large populations: Statistical and methodological issues. *NeuroImage*, 35, 105-120.
- Ullsperger, M., & von Cramon, D. Y. (2003). Error monitoring using external feedback: Specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional Magnetic Resonance Imaging. *The Journal of Neuroscience*, 23, 4308-4314.
- Vanni, S., Tanskanen, T., Seppä, M., Uutela, K., & Hari, R. (2001). Coinciding early activation of the human primary visual cortex and anteromedial cuneus. *Proceedings of the National Academy of Sciences, USA*, 98, 2776-2780.
- Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., Maier, W., Shah, N. J. , Fink, G. R., & Zilles, K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *NeuroImage* 14, 170-181.
- Vrticka, P., Simioni, S., Fornari, E., Schluep, M., Vuilleumier, P., & Sander, D. (2013). Neural substrates of social emotion regulation: A fMRI study on imitation and expressive suppression to dynamic facial signals. *Frontiers in Psychology*, in press.

- Walter, H., Adenzato, M., Ciaramidaro, A., Enrici, I., Pia, L., Bara, B. G. (2004). Understanding intentions in social interactions: The role of the anterior paracingulate cortex. *Journal of Cognitive Neuroscience*, 16, 1854-1863.
- Wolf, I., Dziobek, I., & Heekeren, H.R. (2010). Neural correlates of social cognition in naturalistic settings: A model-free analysis approach. *NeuroImage*, 49, 894-904.
- Worsley K. J., Friston K. J. (1995). Analysis of fMRI time-series revisited – again. *NeuroImage* 2, 173–181.
- Yeung, N. (2013). Conflict monitoring and cognitive control. In K. N. Ochsner & S. Kosslyn (Eds.), *The Oxford Handbook of Cognitive Neuroscience: Volume 2: The cutting edges* (pp. in press). Oxford: Oxford University Press.
- Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp. 1-59). New York: Academic Press.

Figures and Figure Captions

Figure 1

Figure caption

This is how we presented the payoffs in the two states of the world to the sender. Tables 1-3 in the Appendix list all 90 games. Example matrices of the sender-receiver paradigm are given for the three conditions “conflict” (panel A), “sender indifferent” (panel B), and “aligned interest” (panel C). The sender is shown a specific payoff matrix and can send either of two messages: “Red is more profitable for you.” Or “Blue is more profitable for you.” After response selection and on the next screen, the participant has to answer the following question: “Which state do you expect the receiver to choose? The red column or the blue column?” Importantly, the sender’s message does not have a direct impact on the payoffs for both players in any of the states. Rather, the receiver’s choice is eventually implemented for payment.

Example matrices of the sender-receiver paradigm

A

	A	B
Myself	15	5
Player-2	5	15

B

	A	B
Myself	11	11
Player-2	15	10

C

	A	B
Myself	5	6
Player-2	5	15

Figure 2

Figure Caption

Upper Panel: Intention to deceive in strategic interactions: Results are shown for the contrast simple deception *and* sophisticated deception trials versus truth trials.

Lower Panel: Telling the truth: Results are shown for the contrast truth trials versus simple deception *and* sophisticated deception trials.

Abbreviations: aFG: anterior frontal gyrus; MFG: middle frontal gyrus; rTPJ: right temporo-parietal junction. For visualization, a threshold of 99.4% was applied to the probability maps.

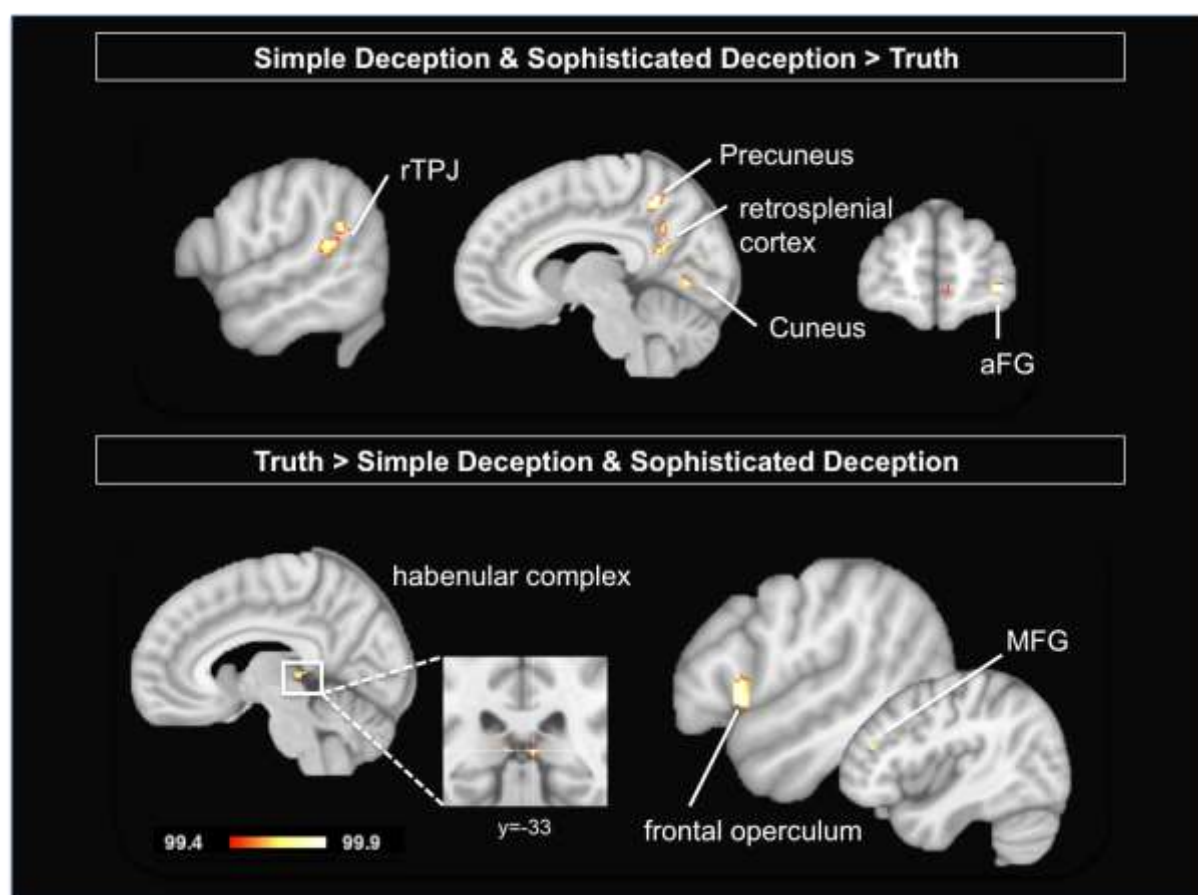


Figure 3

Figure Caption

Upper Panel: Simple Deception: Results are shown for the contrast simple deception trials versus truth trials.

Lower Panel: Sophisticated Deception: Results are shown for the contrast sophisticated deception trials versus truth trials.

Abbreviations: aFG: anterior frontal gyrus; dACC: dorsal anterior cingulate cortex; rTPJ: right temporo-parietal junction. For visualization, a threshold of 99.4% was applied to the probability maps.

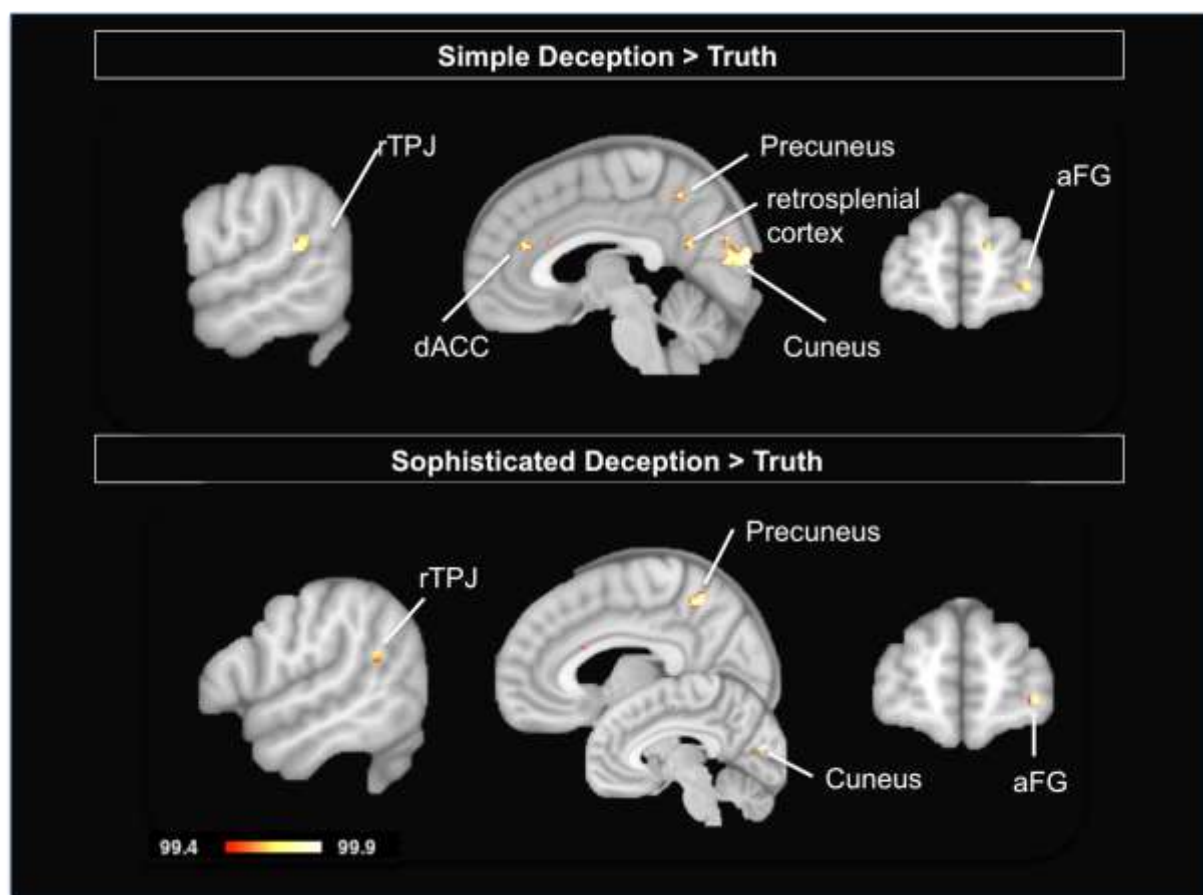


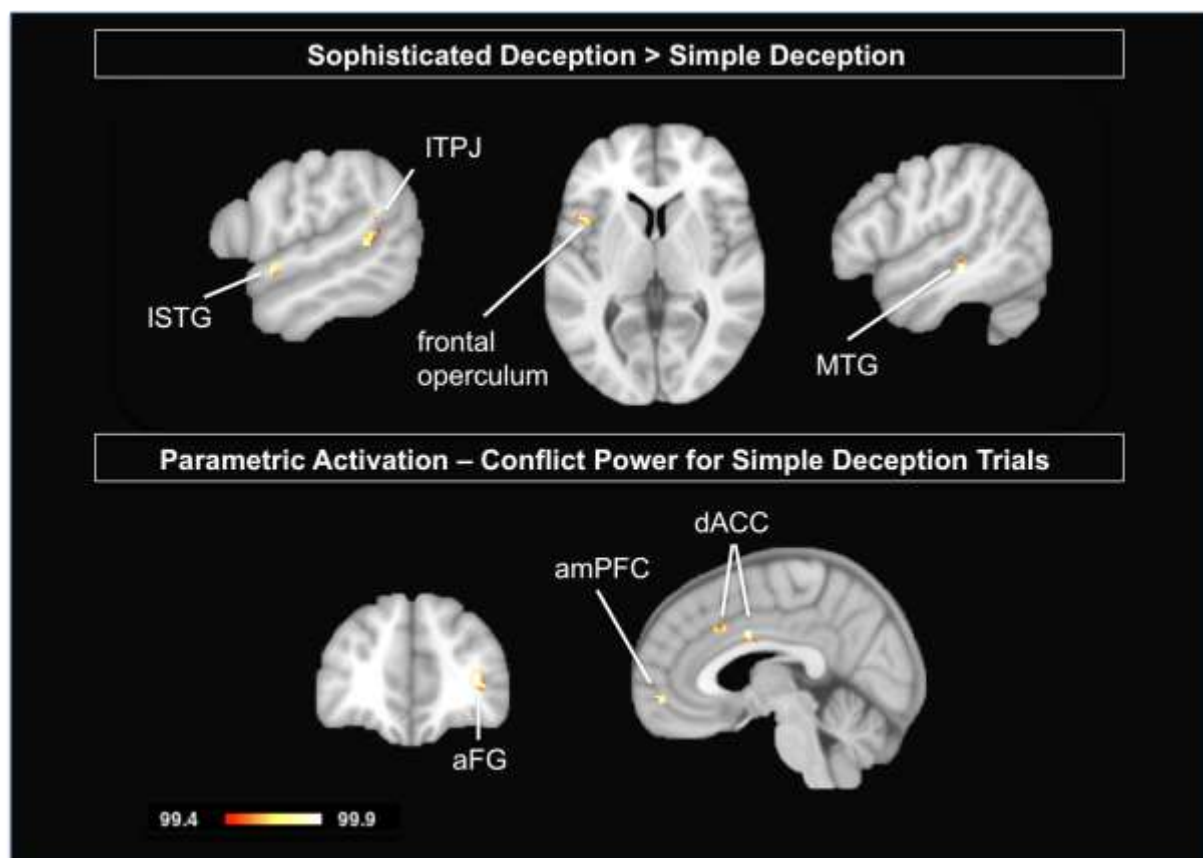
Figure 4

Figure Caption

Upper Panel: Delineating the two forms of deception: Results are shown for the contrast sophisticated deception trials versus simple deception trials.

Lower Panel: Parametric analysis modeling the incentive to deceive for simple deception trials: Results are shown for the positive correlational analysis, i.e., the activation is stronger the higher the conflict and thus the tension in payoffs between sender and receiver.

Abbreviations: aFG: anterior frontal gyrus; amPFC: anterior median prefrontal cortex; dACC: dorsal anterior cingulate cortex; ISTG: left superior temporal gyrus; ITPJ: left temporo-parietal junction; MTG: middle temporal gyrus. For visualization, a threshold of 99.4% was applied to the probability maps.



Tables

Table 1: Performance refers to the answer to the first question: “Which option (blue or red) is more profitable for Player 2?”; intention to deceive refers to the answer to the second question: “Which state do you expect the receiver to choose? The red column or the blue column?”

Performance: Honest answer?	Intention: Intention to deceive?	Trial Classification
Yes	No	Plain truth
	Yes	Sophisticated Deception
No	Yes	Simple Deception
	No	Not classifiable (ignored)

Table 2: Reaction times (in ms) split by category (“aligned interest”, “sender indifferent”, and “conflict”, please cp. section on stimuli and experimental paradigm for more details) and deceptive behavior (truth, sophisticated deception (SD), and plain lies).

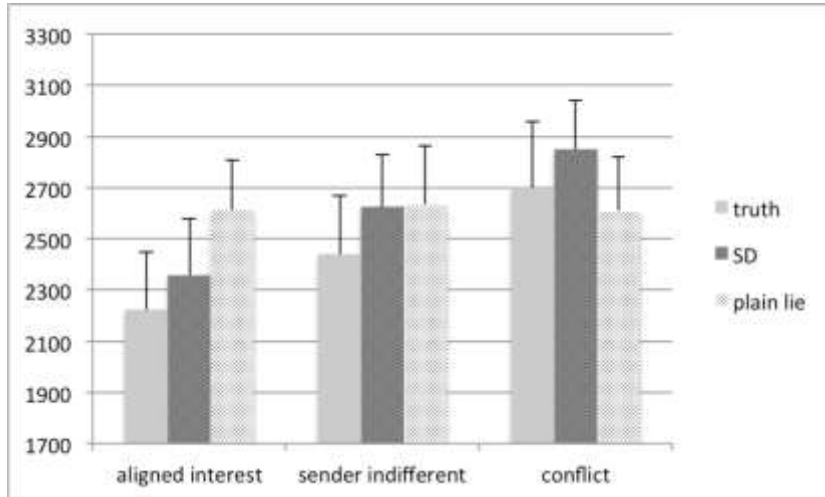


Table 3: Intention to deceive in strategic interactions: Laterality, anatomical specification, Talairach coordinates (x, y, z), posterior probabilities, and size (mm³) for activations according to Bayesian analysis are shown for the contrast simple deception *and* sophisticated deception trials versus truth trials.

<i>Brain region</i>	<i>X</i>	<i>Y</i>	<i>Z</i>	<i>Max</i>	<i>mm³</i>
R. Temporo-parietal junction (TPJ)	55	-42	17	99.92	108
R. Superior temporal gyrus	43	-27	6	99.99	270
R. Precuneus	6	-51	48	99.99	648
Extending into the Retrosplenial cortex	6	-57	20	99.97	189
R. Cuneus	6	-72	-2	99.97	162
	-9	-81	15	99.99	783
	-3	-69	9	99.96	189
R. Superior frontal gyrus (BA 10)	35	57	-2	99.99	216

Table 4: Simple deception versus truth: Laterality, anatomical specification, Talairach coordinates (x, y, z), posterior probabilities, and size (mm³) for activations according to Bayesian analysis are shown for the contrast simple deception trials versus truth trials.

<i>Brain region</i>	<i>X</i>	<i>Y</i>	<i>Z</i>	<i>Max</i>	<i>mm³</i>
R. Temporo-parietal junction (TPJ)	58	-42	17	99.98	648
R. Anterior cingulate cortex (ACC)	3	36	23	99.87	162
R. Precuneus	6	-54	48	99.99	540
Extending into the Retrosplenial cortex	6	-60	23	99.99	
R. Cuneus	6	-90	15	99.99	6021
R. Superior frontal gyrus (BA 10)	14	60	17	99.96	243
	35	57	-2	99.86	162
R. anterior median prefrontal cortex (amPFC)	6	54	4	99.87	108

Table 5: Sophisticated deception versus truth: Laterality, anatomical specification, Talairach coordinates (x, y, z), posterior probabilities, and size (mm³) for activations according to Bayesian analysis are shown for the contrast sophisticated deception trials versus truth trials.

<i>Brain region</i>	<i>X</i>	<i>Y</i>	<i>Z</i>	<i>Max</i>	<i>mm³</i>
R. Temporo-parietal junction (TPJ)	55	-51	23	99.86	162
R. Precuneus	6	-54	50	99.99	1188
L. Cuneus	-6	-81	15	99.99	1107
R. Superior frontal gyrus (BA 10)	32	57	1	99.91	216
R. Superior temporal gyrus	43	-27	6	99.86	162

Table 6: Sophisticated deception versus simple deception: Laterality, anatomical specification, Talairach coordinates (x, y, z), posterior probabilities, and size (mm³) for activations according to Bayesian analysis are shown for the contrast sophisticated deception trials versus simple deception trials.

<i>Brain region</i>	<i>X</i>	<i>Y</i>	<i>Z</i>	<i>Max</i>	<i>mm³</i>
R. Temporo-parietal junction (TPJ)	43	-60	12	99.82	189
L.	-55	-48	12	99.90	270
R. Middle temporal gyrus (MTG)	49	-27	-7	99.97	432
L. Superior temporal gyrus (STG)	-55	0	-2	99.90	243
L. Insula	-40	10	6	99.92	432
R. Mid-cingulate gyrus	6	0	42	99.98	432

Table 7: Truth versus simple and sophisticated deception: Laterality, anatomical specification, Talairach coordinates (x, y, z), posterior probabilities, and size (mm³) for activations according to Bayesian analysis are shown for the contrast truth trials versus simple deception *and* sophisticated deception trials.

<i>Brain region</i>	<i>X</i>	<i>Y</i>	<i>Z</i>	<i>Max</i>	<i>mm³</i>
R. Habenular complex	6	-33	6	99.97	81
L.	-9	-30	6	99.86	108
R. Operculum	49	15	4	99.98	351
L. Pregenuar anterior cingulate cortex	-3	33	4	99.96	135
R. Middle frontal gyrus	41	36	20	99.92	108

Table 8: Parametric analysis modeling the incentive to deceive for simple deception trials: Laterality, anatomical specification, Talairach coordinates (x, y, z), posterior probabilities, and size (mm³) for activations according to Bayesian analysis are shown for the parametric contrast modeling the tension between the sender's and receiver's payoff in simple deception trials.

<i>Brain region</i>	<i>X</i>	<i>Y</i>	<i>Z</i>	<i>Max</i>	<i>mm³</i>
R. anterior median prefrontal cortex (amPFC)	3	54	-5	99.96	297
R. Anterior cingulate cortex (ACC)	6	15	37	99.96	459
R. Middle frontal gyrus (BA 10)	35	42	12	99.97	378

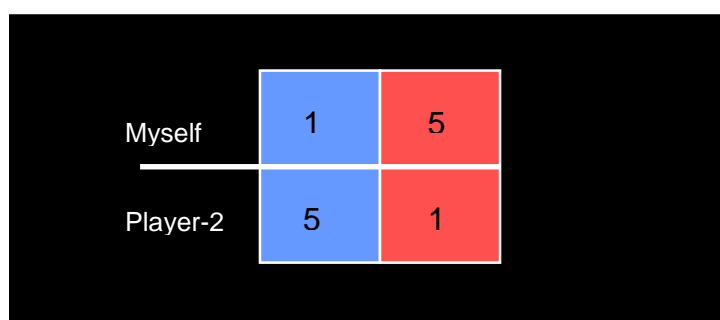
Appendix

A.1 Instructions for the sender (being scanned)

Dear Participant!

Today, you are going to participate in the following experiment:

On the screen you will be presented with a table that might look like this one:



Myself	1	5
Player-2	5	1

Player-2 cannot see the numbers in this table, which correspond to Euros. Player-2 will have to choose either the red or blue column for implementation.

Before Player-2 does so, you can send a message to Player-2 that states which of the columns (**blue** or **red**) is more profitable for Player-2. Note that Player-2 does not get any other information except your message.

If Player-2 decides to choose the **blue** column, you will receive the amount of Euros that are indicated in the **upper blue box** (i.e., 1 Euro in this example) and Player-2 will get the amount of Euros that are indicated in the **lower blue box** (i.e., 5 Euros in this example).

If Player-2 decides to choose the **red** column, you will receive the amount of Euros that are indicated in the **upper red box** (i.e., 5 Euros in this example) and Player-2 will get the amount of Euros that are indicated in the **lower red box** (i.e., 1 Euro in this example).

Thus, you will always get one of the amounts that are displayed in the upper half of the table and Player-2 will always get one of the amounts displayed in the lower half of the table. The choice of Player-2 will determine whether the red or the blue column is relevant for payment.

An Example: You are sending a message to Player-2 that the blue column is more profitable for Player-2. Consider Player-2 indeed picks the blue column. As a result, you would receive 1 Euro and Player-2 would receive 5 Euros.

Procedure

You have two buttons to indicate which message you would like to send:



If you would like to send the message “The **blue** column is more profitable for you”, then press the **blue** button.

If you would like to send the message “The **red** column is more profitable for you”, then press the **red** button.

Please note again, Player-2 does not learn about the specific amounts, but can only choose between the blue and the red column.

After each decision, Player-2 gets the information how many Euros he or she has received in the specific trial. Yet, Player-2 does not learn how much she would have received had she opted for the other column. Neither does she learn about how many Euros you actually had received or how many points you would have received in the other column. Therefore, Player-2 will never be able to judge whether you sent a correct or an incorrect message.

Please note: it is important that you do not ponder for too long for each decisions, but decide within eight seconds, since the response buttons become inactive after eight seconds and the game automatically proceeds. If you accidentally miss one trial, this is not dramatic, but you should try to respond within the eight seconds.

After your choice for sending one of the two messages (and after eight seconds), you will be asked “What do you think which column (blue or red) is Player-2 going to choose?” You can indicate your response again by pressing the left (blue) or right (red) response button. You have four seconds for this response before the experiment automatically proceeds.

All your responses are recorded and after you finished the experiment they will be presented to Player-2 outside the scanner. Being informed about your message for each trial, Player-2 will have to decide on each trial whether to go for blue or red. Player-2 therefore never learns about the allocation of Euros in a particular trial; she only received your message. Thus, Player-2 cannot judge whether your message is correct or incorrect.

By the time when Player-2 finished the experiment, one trial will randomly be selected and paid to both of you accordingly. So, you could get as much as 30 Euros maximum!

If you have questions, please do not hesitate to ask them now, otherwise we start with the training session.

A.2 Instructions for the receiver (recipient of the sender's messages)

Dear Participant!

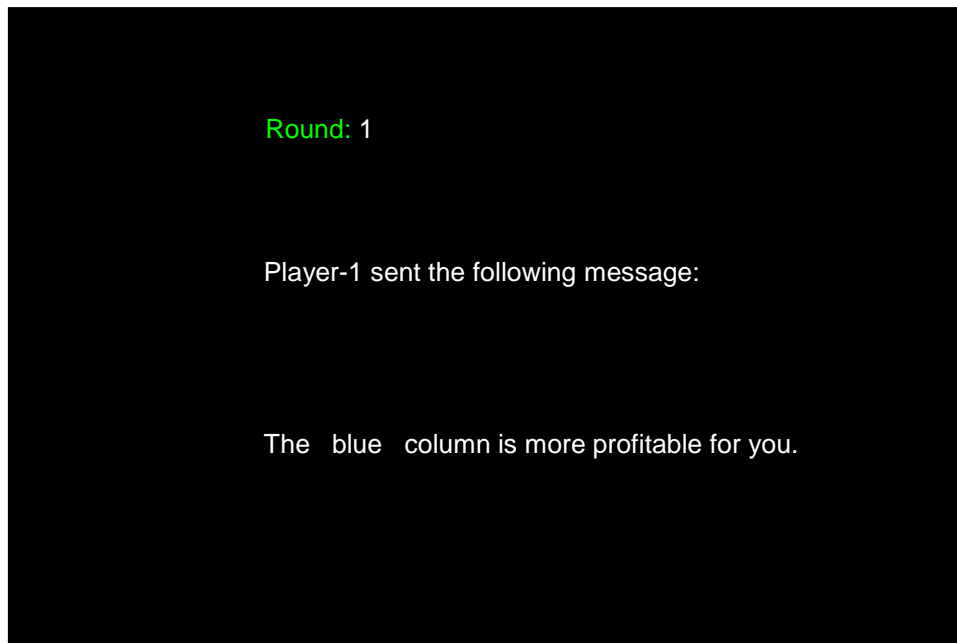
Another player (Player-1) has just finished part 1 of our experiment. In this part, Player-1 was presented with various tables, which differed in the amount of Euros (indicated as numbers in the table) that could be obtained on each trial. Here is an example:

Myself	1	5
Player-2	5	1

The upper half of the table corresponds to the profits Player-1 could get and the lower half to the profits you could get. Hence, you are in the role of Player-2 in this experiment. The profits of both players are linked such that always only ONE column (blue OR red) can be valid while the other becomes obsolete. Which of the two columns is valid entirely depends on YOUR choice, albeit you will never be presented with the specific tables and their associated potential profits.

In contrast, Player-1 is presented with the specific tables and the associated potential profits; yet, Player-1 is not authorized to decide which of the columns will become valid and hence which profits the two of you may receive. However, Player-1 can send you a message whether the blue or the red column is more profitable for you. The messages by Player-1 in each single trial were recorded and stored and will now be loaded accordingly for your decision trials (which we call rounds on the screen).

Here is an example screen:



If you press the blue button, the choice for the blue column is logged in and vice versa for the red button. Once you pressed a button, the next round starts, in which you are asked to make a new decision on which column to choose. Overall, there are 90 trials.

Once more, over the course of the experiment, you will neither learn about the various profits associated with the 90 tables nor about the specific distribution of the profits for you (Player-2) and Player-1 (e.g., identical or opposed profits). Just note, that the maximum profits for you or Player-1 are 30 Euros.

If you decide upon the blue (red) column, the profits in the blue (red) column are paid out to you and Player-1. But note that not all trials are paid out! Once you finished the 90 decisions, only ONE trial is randomly selected and paid to both of you accordingly.

Thank you very much for your attention and enjoy!

A.3 Overview of the full set of games

Listed are all matrices that were employed in the sender-receiver paradigm classified by category. Sender Red: payoff for the sender when state Red is chosen, sender Blue: payoff for the sender when state Blue is chosen, receiver Red: payoff for the receiver when choosing state Red, receiver Blue: payoff for the receiver when choosing state Blue (see also Figure 1 in the manuscript).

Table A1

Listed are all matrices in category “conflict” (n=45).

Sender Red	Receiver Red	Sender Blue	Receiver Blue
1	0	0	1
5	6	6	5
11	10	10	11
16	15	15	16
20	21	21	20
5	0	0	5
5	10	10	5
15	10	10	15
20	15	15	20
20	25	25	20
0	10	10	0
15	5	5	15
10	20	20	10
25	15	15	25
20	30	30	20
0	5	1	0
5	10	6	5
10	15	11	10
16	15	15	20

20	25	21	20
1	0	0	10
5	15	6	5
10	20	11	10
15	25	16	15
21	20	20	30
5	0	0	10
10	5	5	15
15	10	10	20
20	15	15	25
20	30	25	20
0	1	5	0
5	6	10	5
15	10	10	11
20	15	15	16
20	21	25	20
0	1	10	0
15	5	5	6
10	11	20	10
15	16	25	15
30	20	20	21
10	0	0	5
15	5	5	10
10	15	20	10
15	20	25	15
30	20	20	25

Table A2

Listed are all matrices in category “sender indifferent” (n=27).

Sender Red	Receiver Red	Sender Blue	Receiver Blue
1	0	1	1
6	5	6	6
11	10	11	11
1	5	1	0
6	10	6	5
11	15	11	10
1	0	1	10
6	15	6	5
11	20	11	10
5	0	5	5
10	5	10	10
15	10	15	15
5	10	5	0
10	15	10	5
15	20	15	10
5	0	5	15
10	5	10	20
15	10	15	25
10	10	10	0
15	15	15	5
20	20	20	10
10	0	10	15
15	5	15	20
20	25	20	10
10	0	10	20
15	5	15	25
20	30	20	10

Table A3

Listed are all matrices in category “aligned interest” (n=18).

Sender Red	Receiver Red	Sender Blue	Receiver Blue
1	1	0	0
0	0	1	5
1	10	0	0
5	5	6	6
6	10	5	5
5	5	6	15
10	10	11	11
11	15	10	10
10	10	11	20
5	1	0	0
0	0	5	5
5	10	0	0
5	5	10	6
10	10	5	5
5	5	10	15
10	10	15	11
15	15	10	10
15	20	10	10