The neural mechanisms underlying the impact of altruistic outcomes on the process of deception: from the perspectives of communicators and recipients

Dissertation
zur Erlangung der Doktorwürde
der
Philosophischen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität
zu Bonn

vorgelegt von

Lijun Yin
aus
Guangdong, VR. China

Bonn 2017

Gedruckt mit der Genehmigung der Philosophischen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Prof. Dr. Rainer Banse

(Vorsitzender)

Prof. Dr. Martin Reuter

(Betreuer und Gutachter)

Prof. Dr. Bernd Weber

(Gutachter)

Prof. Dr. Henning Boecker

(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 18.01.2017

Acknowledgements

During my Ph.D. journey, I have had ups and downs. I still remember that when I came up with a new idea for an experiment, I was so excited that I could not even fall asleep. But it was difficult to make new experimental ideas come true. Luckily, I have never been alone. I would like to thank Prof. Bernd Weber for providing me the precious opportunity to study in Germany and work on topics that match my interests. His support and supervision mean a lot to me. I would like to thank Prof. Martin Reuter. He gave me precious advice, particularly when I was preparing to publish my first paper and when I was writing my Ph.D. dissertation. I am also grateful to Prof. Li Jian. He provided helpful suggestions during our cooperation on a project and he encouraged me to not give up when we encountered difficulties. I would also like to thank Prof. Zhaoxin Wang, who introduced me to the neuroscience field when I was a master's student and taught me many things that have been valuable to me.

I have had a wonderful time at the CENs. Many thanks to my dear colleagues Yang Hu, Sabrina Strang, Katarina Kuss, Xenia Grote, Markus Antony, Matthias Hampel, Luca Pogoda, Niklas Häusler, Matthias Wibral, Laura Enax, Holger Gerhardt, Heiko Borchersm, Sofia Konrad, Thorben Wölk, Sarah Rudorf, and Hao Chen from the Center for Economics and Neuroscience; Peter Trautner, Amrei König, Vanessa Krapp, Laura Schinabeck, Bettina Mahlow, Karolina Raczka, Marcel Bartling, and Ulrike Broicher from Life & Brain; and Prof. Pia Pinger from the Institute for Applied Microeconomics. They helped me a lot, and it is my pleasure to work with them. I would like to thank all of my friends, particularly Xingnuan Li and Yanjiao Duan, for their company and emotional support during my Ph.D. study. I would also like to thank all the participants who took part in my experiments.

Finally, I would like to express my special thanks to my parents, my sister, my brother in law, and my lovely nephew for their support and love. Without them, nothing would have been possible. I am more thankful than I could ever adequately express. I love them more than anything.

Contents

I. General introduction

Everybody lies and has been lied to. Lying is a widespread phenomenon in society and occurs on a daily basis. On the one hand, deception is treated as a socially unacceptable behavior. When we were kids, we were taught by our parents and teachers that honesty is one of the most important virtues and that we will be punished if we lie. In the famous novel for children "The Adventures of Pinocchio," Pinocchio's lying behaviors are condemned and are punished. As he tells a lie, his nose grows. On the other hand, we were also taught that we should care about others' feelings and welfare and that we sometimes need to lie to preserve others' feelings and welfare. Let us imagine the following: your friend is very ill and his doctor tells you that there is no hope for him to recover. When you are visiting him at the hospital, you probably say to him that he will overcome the illness, although you know he might die soon. Here, lying as a social lubricant can help to ease social interactions, save others from unhappiness, comfort others' misfortune, or even help others.

It is controversial whether lying is acceptable. According to Kantian moral theory, telling a lie is never morally permissible, regardless of the outcomes (Kant, 1797). However, the utilitarianism school of thought implies that lying is morally right if it produces more welfare than any other act could have (Carson, 2010). When you are facing lies and truth in different contexts, how do you choose or feel? What is the neural basis of your decisions to lie and tell the truth, as well as the affections toward lies and truth? How do different outcomes influence your decision whether to lie or tell the truth?

This dissertation is designed to provide some insights into these questions. In the Theoretical section (Section II), the background, the potential impact of altruistic outcomes on the process of deception, and the potential psychological and neural mechanisms underlying deception are introduced. In the Experimental section (Section III), three studies and the corresponding results are reported and discussed in detail. In the General discussion section (Section IV), the findings from the three studies are summarized and discussed.

II. Theoretical section

1. Deception

1.1. Definition of deception

There are many arguments about how deception should be defined. Coleman and Kay emphasized the prototype of lies and three elements of lies: factual falsity, intent to speak falsely, and intent to deceive (Coleman and Kay, 1981). There are two controversial points. First, the factual falsity might not be essential because a communicator might not remember/know the facts and commit an honest mistake (Miller, 1983). It is more important that the communicator believes that his/her statement is false than the factual falsity of the statement. Second, lies do not require the communicator's intention to deceive others, such as the bald-faced lie, which is told by a person who knows that a listener knows it is a lie (Carson, 2010). In this dissertation, the focus is mainly placed on the essence of deception, namely, a deliberate statement (not necessarily an oral one) that is made with the communicator's belief in its falsity, but without providing any clues regarding its falsity[1]. When people talk about deception in their daily lives, they tend to use "lies," "lying," or "deceit" which typically refers to deception in oral or written communication. Therefore, the words "lies," "lying," "deceit," and "deception" are used in turn. Based on the previous definitions (Masip *et al.*, 2004; Alexander, 2007), in this thesis the definition of deception is: "the deliberate or intentional manipulation of information, whether or not it is successful, through the use of verbal or nonverbal cues that the communicator believes are false."

[1] A statement is not a lie, if someone provides the statement which he/she believes it is false and he/she also provides some clues of its falsity, such as "a 'signal of irony'-perhaps by means of tones and choice of words" (Chisholm and Feehan, 1977).

1.2. Categories of lies based on different financial outcomes

In many behavioral experiments about deception, incentivized contexts were used to induce lying behaviors. Based on the financial outcomes of the lies (Gneezy, 2005; Erat and Gneezy, 2012), lies can be classified as *Pareto white lies*, *altruistic white lies*, *selfish black lies,* and *spiteful black lies* (Figure 1.1; the receiver in the figure is the recipient, and the sender is the communicator). *Pareto white lies* increase the payoffs for both the communicator and the recipient. *Altruistic white lies* increase the recipient's payoff at the expense of the communicator's payoff. *Selfish black lies* increase the communicator's payoff at the expense of the recipient's payoff, and *spiteful black lies* decrease the payoffs for both the communicator and the recipient. Here, the difference between Pareto white lies and altruistic white lies is whether a liar sacrifices his/her interests. Similar to this notion, altruistic lies can be defined as "false statements that are costly for the liar and are made with the intention of misleading and benefitting a target" and are classified as a subset of prosocial lies (i.e., "false statements made with the intention of misleading and benefitting a target") (Levine and Schweitzer, 2014).

In this dissertation, prosocial and altruistic lies are not distinguished. The focus is placed on the financial outcomes of certain acts[2] (i.e., lying or truth-telling): 1) other-profit outcomes (altruistic outcomes): the outcomes of lies and truth that financially benefit others (they could also financially benefit the communicators at the same time)[3] and 2) communicator-profit outcomes: the outcomes of lies and truth that only financially benefit the communicators.

[2] There are many other ways of classifying different types of lies. For example, lies can be classified as spontaneous-isolated lies and memorized-scenario lies (Ganis *et al.*, 2003). Nevertheless, other classifications of lies are beyond the scope of the dissertation.

[3] If lies with other-profit outcomes financially benefit both the communicators and the recipients, this type of lies belongs to "Pareto white lies" and "prosocial lies" as mentioned previously. If lies with other-profit outcomes only financially benefit the recipients, this type of lies belongs to "prosocial lies" as mentioned previously.

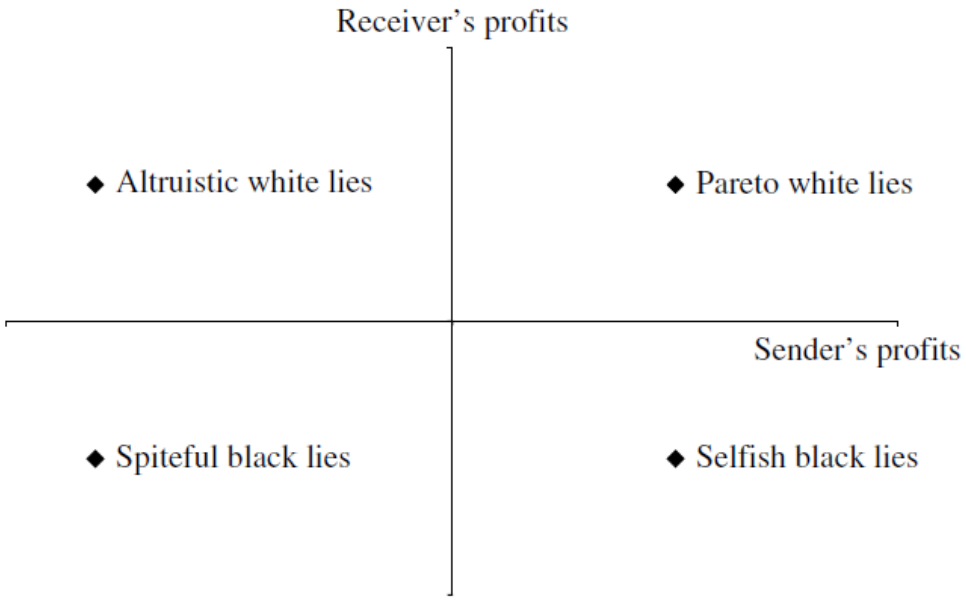

Figure 1.1: Taxonomy of lies based on financial consequences (adapted from Erat and Gneezy, 2012). The origin represents the payoffs of truth-telling. If dots locate above the zero line in the horizontal dimension, receivers' payoffs are increased when senders lie. If dots locate below the zero line in the horizontal dimension, receivers' payoffs are decreased when senders lie. Here, the receiver is the recipient, and the sender is the communicator.

1.3. Psychological costs of lying

Some economic theories suggest that (dis)honest decisions are only determined by the trade-off between punishments if the misreporting is caught and the monetary benefits of successful misreporting (Becker, 1968). Under this assumption, in the absence of punishments or the possibility of being detected, dishonest behaviors should be observed if the behaviors lead to larger monetary benefits. In an incentivized experiment outside the laboratory (Abeler *et al.*, 2014), experimenters phoned a representative sample in Germany and asked them to play a coin tossing game. In the game, reporting tails yielded the payoff of 15€, whereas reporting heads yielded nothing. In this setting, the actual results of coin tossing were only known by the participants themselves. Moreover, lying behaviors were not possible to be

detected or punished. Nevertheless, the aggregate reporting behavior of most of the participants closely followed the expected truthful distribution.[4] When it was financially beneficial to lie, people did not lie maximally (Rosenbaum *et al.*, 2014). The reluctance to deceive might be due to an aversion to lying (Gneezy, 2005; Gneezy et al., 2013) or guilt (Ellingsen et al., 2010; Battigalli et al., 2013), the intention to protect one's positive self-concept or self-image (Mazar and Ariely, 2006; Mazar *et al.*, 2008; Shalvi *et al.*, 2010; Ploner and Regner, 2013), etc.

Accumulating evidence suggests that people have psychological costs of lying (Rosenbaum *et al.*, 2014). In a typical cheap talk sender-receiver game (Gneezy, 2005), private information about two payoff allocations for two players is provided to one of the players as a message sender (Figure 1.2). The other player as a message receiver has no information about the allocations. Each allocation contains a payoff for the sender and a payoff for the receiver. Whenever an allocation is advantageous to the sender, it is disadvantageous to the receiver, and vice versa. The message sender transmits the receiver a message, indicating the advantageous allocation for the receiver. Purely based on the message, the receiver chooses to implement one of the payoff allocations. By manipulating the message, the sender can influence the receiver's decisions. Gneezy invited participants to play the sender-receiver game as senders. In the example presented in Figure 1.2, if a receiver chooses option A, the sender earns $5, and the receiver earns $15. If the receiver chooses option B, the sender earns $15, and the receiver earns $5. Therefore, option A is the advantageous allocation for the receiver and option B is the advantageous allocation for the sender. The sender transmits one of the two messages. Sending message A (i.e., "option A will earn you more money than option B") is truth-telling, whereas sending message B is lying. In addition to the sender-receiver game, Gneezy invited participants to play a dictator game (i.e., a control game) as dictators to investigate participants' decisions

[4] The results might be caused by participants' concerns for anonymity. Some participants might believe that the caller knew their names or address.

in the context without honesty concerns.[5] He found that participants chose the allocations that were advantageous to them more often in the dictator game than in the deception game.[6] In other words, individuals were less willing to earn more by lying. Even in the context where lies helped both the senders and the receivers, a large fraction of the participants still behaved honestly (Erat and Gneezy, 2012). In addition, people tended to avoid settings that enabled them to deceive others (Shalvi *et al.*, 2011b) and even showed a dislike for lies, independent of the outcomes (López-Pérez and Spiegelman, 2013). These results strongly support the notion that people have psychological costs of lying to some extent. Therefore, in addition to the function of an external cost-benefit analysis, the intrinsic costs of lying are important in the deception decision-making process.

1.4. The impact of altruistic outcomes on the process of deception

Batson and Shaw (1991) defined altruism as "… a motivational state with the ultimate goal of increasing another's welfare." By modulating the psychological costs of lying, the altruistic outcomes of lying influence the deceptive decision-making process and the attitude toward lies. For example, altruistic lies might be more acceptable, since they might be good for maintaining a positive self-image or reducing the negative feelings (e.g., guilt or aversion) caused by lying. From the perspective of recipients and third-party observers, studies found that altruistic lies, which benefited others, were judged to be morally appropriate (Hayashi *et al.*, 2014; Levine and Schweitzer, 2014). When making judgments about whether to trust someone, children valued

[5] In a dictator game (Kahneman *et al.*, 1986), a participant as a dictator can choose one of the two monetary allocations (each contains a payoff for the dictator and a payoff for a recipient) to be implemented. The recipient has to passively accept the allocation chosen by the dictator. In this game, the dictator's concerns for the recipient's payoff can be measured. In contrast to the deception game, participants in the dictator game do not need to lie to opponents to earn more money. Therefore, the dictator game is the situation without honesty concerns.

[6] In a deception game, if a sender intends to send a truthful message, he/she should phrase a message by choosing the allocation that is advantageous to a receiver. A message would be untruthful and misleading if it indicates that the disadvantageous allocation to the receiver is advantageous to the receiver (i.e., choosing the allocation that is advantageous to the sender).

both honesty and benevolence (Xu *et al.*, 2013). Individuals who lied to promote the interests of others earned more children's trust than individuals who lied to promote their own interests (Fu *et al.*, 2015). In addition to children, the altruistic outcomes of deception increased trust in adults as well (Levine and Schweitzer, 2015).

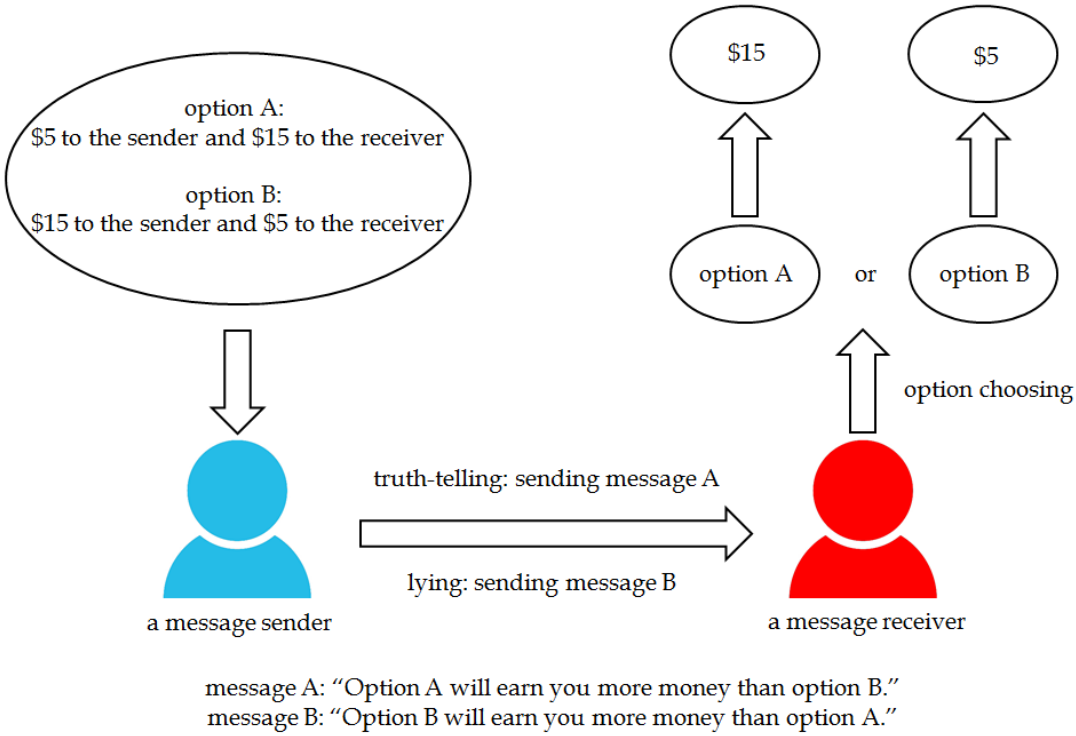

Figure 1.2: The sender-receiver game used in the study by Gneezy (2005). Private information about two payoff allocations (option A and B) for two players is presented to one of the two players as a message sender, whereas the other player as a message receiver has no information about the allocations. Each allocation contains a payoff for the sender and a payoff for the receiver. The sender sends one of the two messages to the receiver (message A or B). After receiving the message, the receiver implements one of the payoff allocations. In this example, the receiver earns more money if he/she implements option A than if he/she implements option B. Therefore, message A is truthful, whereas message B is untruthful.

From the perspective of communicators, dishonesty is more likely to be viewed as morally acceptable and elicits less guilt if it benefits others (Gino *et al.*, 2013). Participants were more likely to cheat when the benefits of cheating were split with another person (Wiltermuth, 2011). In a revised sender-receiver game (Erat and Gneezy, 2012), a message sender has the private information about two allocations and the result of the die roll, which represents one of the two allocations. In the example shown in Figure 1.3, the roll of a die "5" represents option A, which contains a payoff of $20 for the sender and a payoff of $20 for the receiver. The message sender sends a message to a message receiver, indicating the result of the die roll. After receiving the message, the receiver chooses a number. If the receiver chooses the actual outcome of the die roll, the option represented by the outcome will be implemented (i.e., option A). Otherwise, the other option will be implemented (i.e., option B, which contains a payoff of $19 for the sender and a payoff of $30 for the receiver). Erat and Gneezy invited participants to play the game as senders. They found that a significant fraction of the participants (33%; 33 out of 101) told an altruistic lie to profit the receivers, even at the expense of the participants' payoffs. In a die-under-cup game (Shalvi *et al.*, 2011a; Fischbacher and Föllmi-Heusi, 2013), participants were asked to shake a cup to roll a die inside, check the results through the hole at the top of the cup, and report the result of the die roll. The numbers on the die led to different payoffs. When the payoffs would be donated to a cancer charity, 9% of the participants lied that they had rolled a '6', which led to the highest payoff (Lewis *et al.*, 2012). The 9% of the participants is considerably higher than the 2.5% of the participants who lied for participants' personal gains in the study by Shalvi et al. (Shalvi *et al.*, 2011a). The findings from these behavioral studies support that altruistic outcomes influence the process of deception.

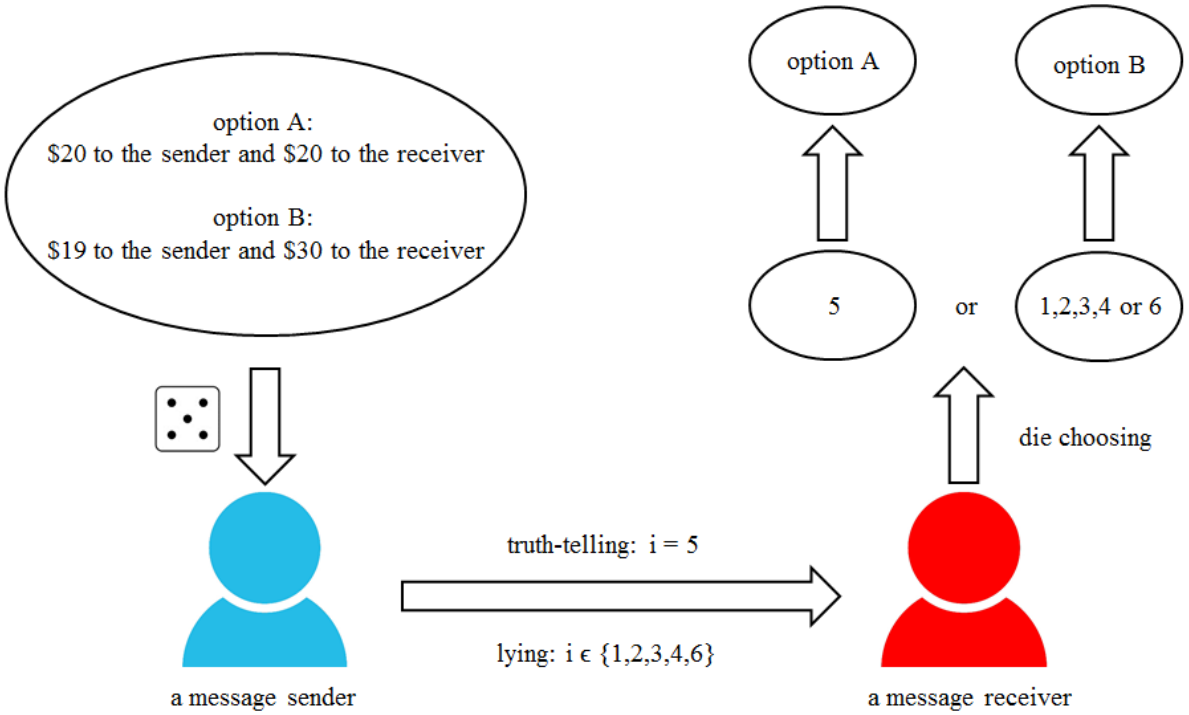

Figure 1.3: The revised sender-receiver game used in the study by Erat and Gneezy (2012). A message sender has private information about two allocations (option A and B), as well as the result of the die roll, which represents one of two allocations (e.g. "5" represents option A). The message sender sends one of the messages to a message receiver. After receiving the message, the receiver chooses a number. If the receiver chooses the actual outcome of the die roll ("5" in this example), option A will be implemented. If the receiver chooses one of the other numbers ("1," "2," "3," "4," or "6" in this example), option B will be implemented. In this example, the message of "The outcome of the roll of die was 5" is truthful.

2. The psychological mechanisms underlying deception

Lying requires multiple psychological processes, including, but not limited to, emotion, executive control, memory, and response inhibition (Spence, 2004; Spence *et al.*, 2004; Vrij, 2008; Farah *et al.*, 2014). Two aspects, cognitive control and emotion, are introduced to provide the readers with some basic knowledge.

2.1. Cognitive control

Predominantly, telling a lie is thought to be more cognitively demanding than telling the truth (DePaulo *et al.*, 2003; Vrij *et al.*, 2006). When an individual tells a lie, the reaction time was longer than the reaction time when he/she tells the truth (Walczyk *et al.*, 2003; Abe *et al.*, 2008). Moreover, liars who reported stories in the reverse order thought they performed worse and behaved more nervously and were easier to be detected as liars by observers than the liars who reported stories in the chronological order (Vrij *et al.*, 2008).

However, the reaction times for lying can be reduced by training or preparation. Compared with participants who were not instructed to speed up lying or were not provided the chance to prepare, participants who were allowed to do so spent less time on providing dishonest responses (Walczyk *et al.*, 2009; Hu *et al.*, 2012). In fact, the differences in the reaction times between lying and truth-telling were erased in participants who received training to speed up their deceptive responses (Hu *et al.*, 2012). Moreover, the high proportion of lying responses made lying easier (Verschuere *et al.*, 2011), meaning that it is easier to lie if one lies more.

Accumulating evidence suggests that honesty also requires considerable time and cognitive control. In the die-under-cup game, the investigators manipulated the time available for participants to report the die roll results. Participants lied less in the situation without time pressure than in the situation with high time pressure (Shalvi

et al., 2012). The result suggests that individuals need sufficient time to enhance the level of self-control and resist the temptation to lie for more benefits (Rosenbaum *et al.*, 2014). Furthermore, when participants' self-control resources were depleted, the dishonest behaviors increased (Mead *et al.*, 2009; Gino *et al.*, 2011). In the afternoon, when the self-control resources were more intensely depleted, participants lied or cheated more often than in the morning (Kouchaki and Smith, 2014). These findings suggest that considerable time and cognitive control are also important for making honest decisions.

2.2. Emotion

Emotion has been long emphasized to have an important role in moral behavior (Eisenberg, 2000). Higher-order emotions, such as guilt and shame (with negative valence), and basic emotions, such as anger and fearfulness, arise from the violation of a moral standard or moral transgression (Tangney *et al.*, 1992; Eisenberg, 2000). Emotional processes are also engaged in moral judgment to ground (Greene *et al.*, 2001; Valdesolo and DeSteno, 2006), rise from, and possibly further drive or modulate moral decision-making (Tangney *et al.*, 2007; Malti and Krettenauer, 2013). More specifically, by influencing approach-avoidance tendencies, such as achieving anticipated positive emotions and avoiding anticipated negative emotions, individuals might be motivated to make different moral decisions (Baumeister *et al.*, 2007).

As moral behaviors, lying and truth-telling, have a strong association with emotions (Farah *et al.*, 2014). Liars have strong emotional experiences, such as guilt or fear (Ekman, 1985, 1989). The emotion experienced when an individual decides to lie and the expectation of the emotional experience of lying might influence deception-related decisions (Gaspar and Schweitzer, 2013).

3. The neural mechanisms underlying deception

3.1. Functional magnetic resonance imaging

During the last decade, the use of functional magnetic resonance imaging (fMRI) to investigate the neural correlates of deception has increased dramatically. This technique measures brain activity by detecting the changes associated with blood flow, which supports increased neuronal activity (Huettel *et al.*, 2004). The following brief introduction of this technique and data analysis methods is based on multiple resources (Ogawa *et al.*, 1990; Huettel *et al.*, 2004; Amaro and Barker, 2006; Poldrack, 2007; Poldrack *et al.*, 2011; Lindquist and Wager, 2014).

3.1.1. A brief introduction of the fMRI technique

When neurons in the brain become active, the amount of blood flowing through the area is increased. More oxygen is used by the cells and further leads to a relative increase in the local blood oxygen levels. The changes in the state of hemoglobin oxygenation can be detected by blood oxygenation level-dependent (BOLD) contrast images. The oxygen concentration endows the hemoglobin molecule with different magnetic properties. Oxyhemoglobin is diamagnetic, whereas deoxyhemoglobin is paramagnetic, which distorts the magnetic field. The diamagnetic blood interferes with the MR signal less, and, therefore, the areas with higher oxyhemoglobin concentrations provide a higher signal than the areas with low concentrations. The change in the MR signal from the neuronal activity is the hemodynamic response (HDR). One to two seconds after the neuronal events, the BOLD signal begins and increases to a peak at approximately five seconds. After the neurons stop firing, the

BOLD signal falls below the original level (i.e. undershoot[7]), and the signal recovers to the baseline over time.

3.1.2. Preprocessing steps

The aim of fMRI data analyses is to estimate the changes in the BOLD signal in response to some stimulation or manipulation. There are some basic preprocessing steps during the initial analysis of the fMRI data. These steps usually include slice timing correction, motion correction, coregistration, normalization, and spatial smoothing. Quality control of the data is crucial. However, none of these operations is mandatory and necessary in all cases. Here, the basic preprocessing steps are briefly introduced.

The slice timing correction corrects for the acquisition time of each slice of one functional volume. Different slices from one functional volume are acquired sequentially in time. The slices can be acquired in ascending order, descending order or by using interleaved acquisition. A reference slice is usually chosen to correct the mismatch between the acquisition time of different slices. The data in other slices can be interpolated to match the timing of the chosen slice.

Motion correction is performed to overcome the problem induced by participants' movement during the scanning. Due to excessive motion, the voxel's intensity might be contaminated by the signal from the nearby voxels. A reference image (usually the first or the mean image) is chosen to compensate for participants' movement. A rigid body transformation is applied to the other images to match the reference image. Algorithms are used to search for the best parameter estimates to translate and rotate the images to match the reference image. The estimated motion parameters are usually used as covariates in the first level of analysis. Because this strategy is unable

[7] The venous bed capacity makes the regional blood volume normalize more slowly than the changes in blood flow, and leads to high deoxy-hemoglobin concentration (Jones *et al.*, 1998).

to correct for more complex artifacts caused by severe head motion, a participant's data can be excluded if his/her movement is too severe.

Coregistration is used to map the results from the analysis of low-resolution functional images onto a high-resolution structural image. The structural image is usually co-registered to the first or the mean functional image and warps to a standard template. The transformations can be applied to the functional images. This step is also associated with the next phase: normalization. Normalization is used to register the data into a standard anatomical space, which can be reported in standard coordinate systems, including the Montreal Neurologic Institute (MNI) and Talairach systems.

Spatial smoothing is used to blur the image intensities by applying a spatial smoothing kernel. The amount of smoothing is determined by the width of the distribution. If the Gaussian kernel is used, the width of the distribution is the full width at half-maximum (FWHM). By removing the high-frequency information, smoothing improves inter-subject registration and increases the signal-to-noise ratio.

3.1.3. Individual-level analyses and group-level analyses

A general linear model (GLM) is often used to analyze a voxel's time series to determine if its BOLD signal matches the presented pattern of multiple events. A GLM analysis is typically a univariate analysis and is performed for every voxel in a single subject. The BOLD time series at every voxel can be expressed as a linear function of a design matrix, the vector of parameter estimates β, and the error term. The β value is estimated and used to test for an effect of interest. Since the GLM is applied to each voxel across the brain, the effect of interest is to determine the voxels that are statistically significantly activated by the experimental conditions. In the statistical analysis, the null hypothesis (H_0) is that the manipulation or the independent variable in the experiment has no effect, and there is no difference among the β values under different experimental conditions (e.g., H_0: $\beta_1 = \beta_2$). The research hypothesis (H_1) is that the manipulation or the independent variable has an

effect, and there are differences between the β values under the experimental conditions (e.g., H_1: $\beta_1 \neq \beta_2$).

The group-level analyses are used to combine the results from single subjects to obtain group results and test the hypotheses at the group level. One important issue during the group-level analyses is the multiple comparisons correction. Each voxel needs at least one hypothesis test, and, therefore, there are up to thousands of statistical tests that must be performed. Multiple comparisons might result in false positives. Some approaches, such as family-wise error (FWE) and false discovery rate (FDR) corrections, are used to control the false positive rates.

The standard approach to analyze the fMRI data is whole brain analysis, which focuses on the data for each voxel in the whole brain. In addition, region-of-interest (ROI) analyses can be used if the researchers have a specific hypothesis about a specific brain region (ROI) that is associated with a particular cognitive process, rather than the whole brain. The ROIs can be defined using anatomical criteria or functional activity maps. The common approach is to extract the signal from the defined ROIs and perform further analyses. ROI analyses reduce the total number of statistical comparisons, minimize the need for multiple comparison corrections, and increase the signal-to-noise ratio.

3.2. Previous neuroimaging studies of deception

3.2.1. Paradigms used in previous neuroimaging studies of deception

In many previous neuroimaging studies of deception, a typical paradigm called "instructed paradigms" was used to investigate the neural correlates of lying (Greely and Illes, 2007; Sip *et al.*, 2008; Schauer, 2010; Wright *et al.*, 2013; Farah *et al.*, 2014). In this type of paradigm, investigators usually instruct participants to provide truthful or untruthful answers to some questions at a specific time. The contents of the questions included memories (Spence *et al.*, 2001; Abe *et al.*, 2007; Abe *et al.*,

2008; Bhatt *et al.*, 2009; Ito *et al.*, 2011), possession of an item (Langleben *et al.*, 2002; Luan Phan *et al.*, 2005), personal information or experience (Ganis *et al.*, 2003; Nunez *et al.*, 2005; Abe *et al.*, 2006; Ganis *et al.*, 2009; Ganis *et al.*, 2011), knowledge of a mock crime (Mohamed *et al.*, 2006; Kozel *et al.*, 2009a; Kozel *et al.*, 2009b), or valence of pictures (Lee *et al.*, 2010). For example, in a word memory task (Abe *et al.*, 2008), participants were asked to learn words during the study phases and provide responses during the test phrases. During the test phase, participants were placed in the MRI scanner and observed previously studied words ("True targets"), non-studied words ("False targets" that are semantically related to the studied items), and unrelated words ("New targets"). "True targets," "False targets," and "New targets" were assigned to "Truth" blocks, in which participants were asked to provide a truthful response to each stimulus. "True targets" and "New targets" were assigned to "Lie" blocks, in which participants were asked to provide an untruthful response to each stimulus, namely "New" responses for "Truth targets" and "Old" responses for "New targets."

In recent years, another type of experimental paradigm emerged (hereafter called spontaneous paradigms). This type of paradigm allows participants to freely make honest or dishonest decisions by providing motivations to lie, such as monetary rewards (Baumgartner *et al.*, 2009; Greene and Paxton, 2009; Bhatt *et al.*, 2010; Sip *et al.*, 2010; Sip *et al.*, 2012; Baumgartner *et al.*, 2013; Abe and Greene, 2014; Sun *et al.*, 2015a; Volz *et al.*, 2015), or reducing the time of the experiment (Ding *et al.*, 2013). For example, in a coin-flip task (Greene and Paxton, 2009; Abe and Greene, 2014), participants were asked to predict the outcomes of coin flips in the scanner. Every correct prediction led to a monetary benefit. In the no opportunity condition, participants had to explicitly report their prediction and then indicate whether their prediction was correct. In the opportunity condition, participants did not need to explicitly report their prediction. Therefore, they had the chance to win more money

by lying about the correctness of their prediction in the opportunity condition. In this task, participants were able to spontaneously make decisions to lie or to tell the truth.[8]

Although the instructed paradigms are relatively easy to perform and provided us with a large amount of knowledge about the potential neural correlates associated with lying, there are some limitations when researchers use them to investigate deception. For example, the paradigms might have poor ecological validity (Ganis and Keenan, 2009) and different involvements of emotions and cognitive control (Farah *et al.*, 2014). It seems that lies in the spontaneous paradigms might be more similar to real-life lies. Therefore, it is worth investigating the differences in the processes of lying and truth-telling between instructed and spontaneous paradigms.

3.2.2. Brain regions associated with deception

Meta-analyses of deception, which are mostly based on instructed experiments (Christ *et al.*, 2009; Farah *et al.*, 2014), revealed several brain regions that were more active during lying than truth-telling. These regions include the dorsal lateral prefrontal cortex (DLPFC) (Lee *et al.*, 2002; Lee *et al.*, 2005; Luan Phan *et al.*, 2005; Nunez *et al.*, 2005; Abe *et al.*, 2006; Abe *et al.*, 2007), the superior frontal gyrus (SFG) (Langleben *et al.*, 2002; Langleben *et al.*, 2005; Ganis *et al.*, 2009; Lee *et al.*, 2010; Sun *et al.*, 2015b), the inferior frontal gyrus (IFG) (Langleben *et al.*, 2005; Nunez *et al.*, 2005; Christ *et al.*, 2009; Ganis *et al.*, 2009; Lee *et al.*, 2010; Sun *et al.*, 2015b), the insula (Nunez *et al.*, 2005; Mohamed *et al.*, 2006; Christ *et al.*, 2009; Ganis *et al.*, 2009; Kozel *et al.*, 2009a; Lee *et al.*, 2009; Ito *et al.*, 2011), the anterior cingulate cortex (ACC) (Langleben *et al.*, 2002; Ganis *et al.*, 2003; Lee *et al.*, 2005; Mohamed *et al.*, 2006; Christ *et al.*, 2009; Lee *et al.*, 2009), the inferior parietal lobule (IPL) (Spence *et al.*, 2001; Lee *et al.*, 2002; Christ *et al.*, 2009; Lee *et al.*, 2010; Cui *et al.*, 2014; Sun *et al.*, 2015b), and the middle frontal gyrus (MFG) (Ganis *et al.*, 2003;

[8] Despite the fact that in the opportunity condition participants could freely make their decisions to lie or to tell the truth, the experimental design suffered the drawback that the researchers could not identify lying and truth-telling trials because they did not know participants' actual predictions.

Langleben *et al.*, 2005; Nunez *et al.*, 2005; Christ *et al.*, 2009; Ganis *et al.*, 2009; Kozel *et al.*, 2009a; Lee *et al.*, 2009; Lee *et al.*, 2010; Ito *et al.*, 2011; Sun *et al.*, 2015b). The next section presents a brief introduction to the brain regions associated with deception. These areas are also thought to relate to cognitive control, emotion, and reward-processing.

3.2.2.1. Cognitive control

Section 2.1 presents the findings from previous behavioral studies of the reaction times for lying and truth-telling, as well as the influence of self-depletion and time-limit on the lying frequencies. In addition to the behavioral findings, new perspectives are provided by the neuroimaging studies of deception. Attempted lying was linked to the activation of executive brain regions (Christ et al., 2009; Farah et al., 2014). A meta-analysis compared the activation likelihood estimate (ALE)[9] maps of three aspects of executive control (i.e., working memory, inhibitory control, and task switching) with the ALE maps of deceptive responses (versus truthful responses) (Christ *et al.*, 2009). The deception related regions, such as the VLPFC, the anterior insula (AI), and the ACC, were also activated in the tasks associated with the executive control. In addition, the DLPFC and the posterior parietal cortex were overlapping regions that were activated in the tasks associated with both deception and working memory. Therefore, lying was thought to be more cognitively demanding, whereas a truthful response was treated as a default behavior (Spence et al., 2004).

The instructed paradigms were often used in studies supporting the notion of a higher requirement for cognitive control during lying. Nevertheless, in the study by Greene and Paxton (2009), which used a spontaneous paradigm, truth-telling is believed to be cognitively demanding. According to the self-report accuracy in the opportunity

[9] The activation likelihood estimation quantifies the degree of anatomical overlap across neuroimaging studies that are based on peak-voxel coordinate information (Farah *et al.*, 2014).

condition, participants were classified as honest, dishonest or ambiguous.[10] In the comparison of the loss trials in the opportunity condition[11] and the loss trials in the no opportunity condition,[12] no significant effects were identified in honest participants. In contrast to the honest participants, dishonest participants showed activation in the ACC, the DMPFC, the DLPFC, the VLPFC (control-related prefrontal regions), and the right parietal lobe. Moreover, increased activity in the bilateral DLPFC was also identified in the dishonest participants in the comparison of win trials in the opportunity condition[13] and win trials in the no opportunity condition.[14] Since the DLPFC was activated in both the lying associated contrast and the truth-telling associated contrast, the researchers concluded that the DLPFC participates in the process of actively deciding whether to lie, regardless of the choice made. In addition, dishonest individuals might require extra effort when they choose to forgo opportunities for dishonest gain.

Given the potential differences in the neural correlates of spontaneous and instructed lying/truth-telling, a meta-analysis was performed to investigate the different neural processes of lying in instructed studies and volitional studies (Lisofsky *et al.*, 2014). The results showed an increased activation in the bilateral IPL during volitional lying (versus instructed lying). In the volitional studies included in this meta-analysis, participants were not instructed to respond truthfully or untruthfully at a specific time. Nevertheless, in some of the studies, participants were asked to follow some

[10] In Greene and Paxton's experiment (2009), the chance level of the accuracy should be 50%. Fourteen participants with the highest accuracies (higher than 69%, mean = 84%) were classified as dishonest. Fourteen participants with the lowest accuracies (mean = 52%) were classified as honest. Others were classified as ambiguous (N = 7; mean = 62%).
[11] In the loss trials from the opportunity condition, the behavior of claiming that their predictions were wrong, to a great extent, is honest.
[12] In the loss trials from the no opportunity condition, the behavior of claiming that their predictions were wrong is honest.
[13] In the win trials from the opportunity condition, dishonest participants' responses could be either truthful (honest) or untruthful (dishonest). Because the actual predictions were unknown, honest and dishonest responses in the wins trials from the opportunity condition could not be identified in this experiment.
[14] In the win trials from the no opportunity condition, the behavior of claiming that their predictions were right is honest.

rules. For instance, participants should imagine monetary gains from their successful feigning (Lee *et al.*, 2002; Browndyke *et al.*, 2008; Lee *et al.*, 2009; McPherson *et al.*, 2011) or achieve an approximate balance between honest and dishonest responses (Spence *et al.*, 2008).

3.2.2.2. Emotion

Emotion is essential for human behaviors (Pessoa, 2009). Regions such as the amygdala, the insula, the orbitofrontal cortex, and the anterior cingulate cortex are related to different basic emotions (anger, fear, disgust, sadness, and happiness), different dimensions of emotions (arousal and valence) (Sprengelmeyer *et al.*, 1998; LeDoux, 2000; Phelps *et al.*, 2001; Phan *et al.*, 2002; Anders *et al.*, 2008; Hamann, 2012), social emotions (Lamm and Singer, 2010), and moral emotions (Moll *et al.*, 2002b).

From the perspective of communicators, a large amount of neuroimaging literature consistently showed the involvement of the anterior insula and the amygdala in lying (Abe *et al.*, 2007; Baumgartner *et al.*, 2009; Christ *et al.*, 2009; Baumgartner *et al.*, 2013; Farah *et al.*, 2014; Lisofsky *et al.*, 2014; Sun *et al.*, 2015a; Volz *et al.*, 2015). In a study of broken promises (Baumgartner *et al.*, 2009), researchers used a modified version of a trust game, where participants acted as trustees (second movers) and first made a promise about whether they would always, mostly, sometimes or never pay back real money to an investor (first mover). Based on the trustees' promises, opponents acting as investors decided whether to trust the trustees and invest money. Afterward, participants chose to keep or break the promise they made. Compared with honest participants who behaved more trustworthily, dishonest participants showed significant activation in the ACC and the DLPFC in the trials with promises (versus trials without promises) during the first 6 seconds of the decision phase (phase A). In addition, increased activity was observed in the amygdala in the second 3 seconds until the button press (phase B). The authors further found that the return rates negatively correlated with activity in the ACC and the left DLPFC during phase A

and the left amygdala during phase B. These findings suggest that the dishonesty involves an emotional conflict.

From the perspective of a third party or a recipient, emotion-related brain regions were also involved during the process of lying behaviors. In a study that investigated the judgment about others' intention to deceive, participants watched videos of several actors lifting a box and judged if the actors were trying to deceive the observers about the box's weight (Grèzes *et al.*, 2004). The amygdala and the anterior cingulate cortex were activated when participants judged that the actors were misleading the observers. In another study (Grezes *et al.*, 2006), participants watched videos of actors (either themselves or others) lifting a box and judged if the experimenter had misled the actors or if the actors had misled the experimenter as to the actual weight of the box. The amygdala was activated upon the judgment that the experimenter misled participants, compared with the other judgments (participants misled the experimenter, others misled the experimenter, and the experimenter misled others). These findings suggest that activity in the amygdala was modulated by the affective reaction that occurs when detecting deceiving behaviors or being deceived.

3.2.2.3. Reward processing

The motivation to achieve pleasant states and avoid unpleasant states guides human behavior and decisions (Daw et al., 2006; Linke et al., 2010). The striatum, particularly the nucleus accumbens (NAcc; part of the ventral striatum), is a key structure linked to the anticipation of rewards, such as monetary benefits (Delgado *et al.*, 2000; Berns *et al.*, 2001; Knutson *et al.*, 2001a; Knutson *et al.*, 2001b; O'Doherty, 2004; Bartra *et al.*, 2013) or social rewards (Izuma *et al.*, 2008; Spreckelmeyer *et al.*, 2009; Häusler *et al.*, 2015). However, activation in the reward-related regions was rarely observed in previous studies of deception, particularly in the studies that used the instructed paradigms. The results might be due to the lack of monetary or social rewards as the consequences of successful lying. The paradigms that can motivate

people to lie by providing monetary or social rewards are important to better understand the reward process in deception. Abe and Greene (2014) performed a study to investigate the relation between anticipated reward and dishonest behaviors. In addition to the coin-flip task used in the study by Greene and Paxton (2009), participants performed a monetary incentive delay (MID) task,[15] which was used to maximize the affective and motivational aspects of reward processing (Knutson *et al.*, 2000). The authors found that the individuals who had stronger NAcc responses to the anticipated reward in the MID task exhibited 1) higher rates of dishonest behavior (indexed by improbably high levels of self-reported accuracy in the coin-flip task) and 2) greater involvement of the DLPFC when refraining from dishonest gain. The results suggest the important roles of the reward-related regions, particularly the NAcc, in (dis)honest decision-making in the incentivized contexts.

Furthermore, the reward-related regions might also be related to the process or the expectation of dishonest gain. Successful lies also elicited higher activation in the ventral striatum compared with lies that were caught and punished (Sun *et al.*, 2015a). In the broken promises study (Baumgartner *et al.*, 2009), compared with the honest participants, significant activation in the right ventral striatum was observed in dishonest participants in the trials with promises, which might increase the chance of obtaining the dishonest gain. Therefore, deception, particularly in the incentivized contexts, might require the involvement of the reward-related brain regions to process potential rewards.

[15] In the MID task, participants press a button during the brief presentation of a target (duration of the presentation: 0.15-0.45s) to earn a financial reward or to avoid a financial loss. They then get the feedback of their performance.

4. Summary

To date, the neural mechanisms underlying the process of deception are not well understood, particularly the differences between spontaneous and instructed lying and truth-telling, as well as the impact of altruistic outcomes on the process of deception. Three studies were performed and reported in the Experimental section to provide some insights into these questions (Section III). Study 1 was performed to investigate the different neural mechanisms underlying lying and truth-telling between the recently evolved spontaneous paradigms, in which participants make decisions on their own initiative, and the instructed paradigms, in which participants make decisions to lie and tell the truth at a specific time by following the instructions. The aims of Study 2 and Study 3 are to investigate the impact of altruistic outcomes on the neural processes underlying lies and truth. Specifically, Study 2 was performed to investigate the impact of altruistic and self-profiting outcomes on the deceptive decision-making process and the underlying neural mechanisms. Study 3 was performed to investigate the recipients' neural responses to lies and truth based on outcomes that are either beneficial or harmful to the recipients.

III. Experimental section

5. Study 1: Different neural mechanisms underlying spontaneous and instructed deceptive decision-making process

The associated work has been published in Yin Lijun, Reuter Martin and Weber Bernd (2016). Let the man choose what to do: Neural correlates of spontaneous lying and truth-telling. Brain and Cognition, 102, 13-25.

Author contributions: L.Y. and B.W. designed research; L.Y. performed research; L.Y. analyzed data; L.Y., M.R., and B.W. wrote the paper.

5.1. Introduction

Experimental paradigms are crucial to investigate the neural correlates of deception. In the last decade, popular paradigms used to investigate the neural processes involved in lying are called "instructed paradigms" (please see Section 3.2.1 for more details). Since participants are not usually allowed to make their choices freely, the instructed paradigms have been criticized. In particular, the mental processes of lying in the instructed paradigms are thought to be widely different from lies in real life, and, therefore, the instructed paradigms are not suitable for investigating deception (Greely and Illes, 2007; Sip *et al.*, 2008). Unlike the instructed paradigms, the recently developed spontaneous paradigms allow individuals to freely decide whether to lie and when to lie.

The aim of Study 1 is to investigate the neural mechanisms underlying (un)truthful responses, which were given on one's own initiative and by following others' instructions. A modified *sic bo*[16] gambling game (Eadington, 1999) was adopted to

[16] The original *Sic bo* game is a casino game, in which people can bet on one of multiple options based on different combinations of three dice.

allow individuals to be more involved in the experiment. In this game, individuals were asked to predict the outcomes of three dice rolls and bet real money. After the outcomes of the rolls were presented, they reported whether their predictions were correct and they were paid according to their reports. In the spontaneous session, participants could freely make decisions. If their predictions were wrong, they could still win the stakes by providing untruthful responses with regard to their predictions. In the instructed session, free decisions were not allowed. Participants should report their betting results truthfully or untruthfully, according to the presented instructions. When their predictions were wrong, they could only win the stakes if the instructions asked them to provide untruthful responses with regard to their predictions. Otherwise, they would lose the stakes.

5.2. Materials and methods

5.2.1. Participants

Fifty-four healthy male participants[17] (mean ± s.d. age = 26.1 ± 3.8 years ranged from 19 to 36 years) took part in the fMRI experiment. All participants had no history of substance abuse, psychiatric or neurological disorders. They had normal or corrected-to-normal vision. The ethics committee of the Medical Faculty of the University of Bonn approved the study. All participants provided their written informed consent.

5.2.2. Tasks

Study 1 used a modified *sic bo* game (Figure 5.1). In the original game, combinations of three dice formed various betting options. In the simplified version used here, only two betting options were available ("big" if the sum of the three dice is from 11 to 18

[17] Some studies found that male and female tend to tell different types of lies. Men told more self-oriented lies than women, whereas women told more other-oriented lies than men (DePaulo *et al.*, 1996; Feldman *et al.*, 2002). In Study 1, male participants were recruited, in order to enlarge the sample size of the partially dishonest group and to avoid potential gender differences in the process of lying.

and "small" if the sum of the dice is from 3 to 10)[18]. In each trial, participants were given a budget of 25€. At the beginning of each trial, they should predict the sum of the three dice and bet on either "big" or "small" by pressing the button on the response grips within 2.5s. After that, one image of three dice was presented, and participants knew whether their prediction was correct. It was followed by a fixation cross for 2-6s. A certain stake (see Section 5.2.3 for more details) was then presented, accompany with a question that whether their prediction was correct. Participants should answer the question by pressing the button on the response grips within 3.5s. If participants chose "Yes", they would win the stake. Otherwise, they would lose the stake. In the first session of the experiment (i.e., spontaneous session/condition; Figure 5.1*A*), participants could freely make decisions to lie or to tell the truth. In the second session of the experiment (i.e., instructed session/condition; Figure 5.1*B*), participants would first see the instruction (i.e., "Right answer" or "Wrong answer"). According to the instruction, they provided truthful or untruthful answers. To be more specific, if the instruction was "Wrong answer" (i.e., instructed lying), participants should choose "No" in the trials where their predictions were correct and choose "Yes" in the trials where their predictions were incorrect. The positions of the betting options and the report options were counterbalanced within each session and each participant.

[18] In the original gambling game, when the sum is from 4 to 10, gamblers win if they bet on "small". But the 3 (3 ones) is not a winning bet in this case. When the sum is from 11 to 17, gamblers win if they bet on "big". But the 18 (3 sixes) is not a winning bet in this case. To simplify the game, the rules have been changed as described here.

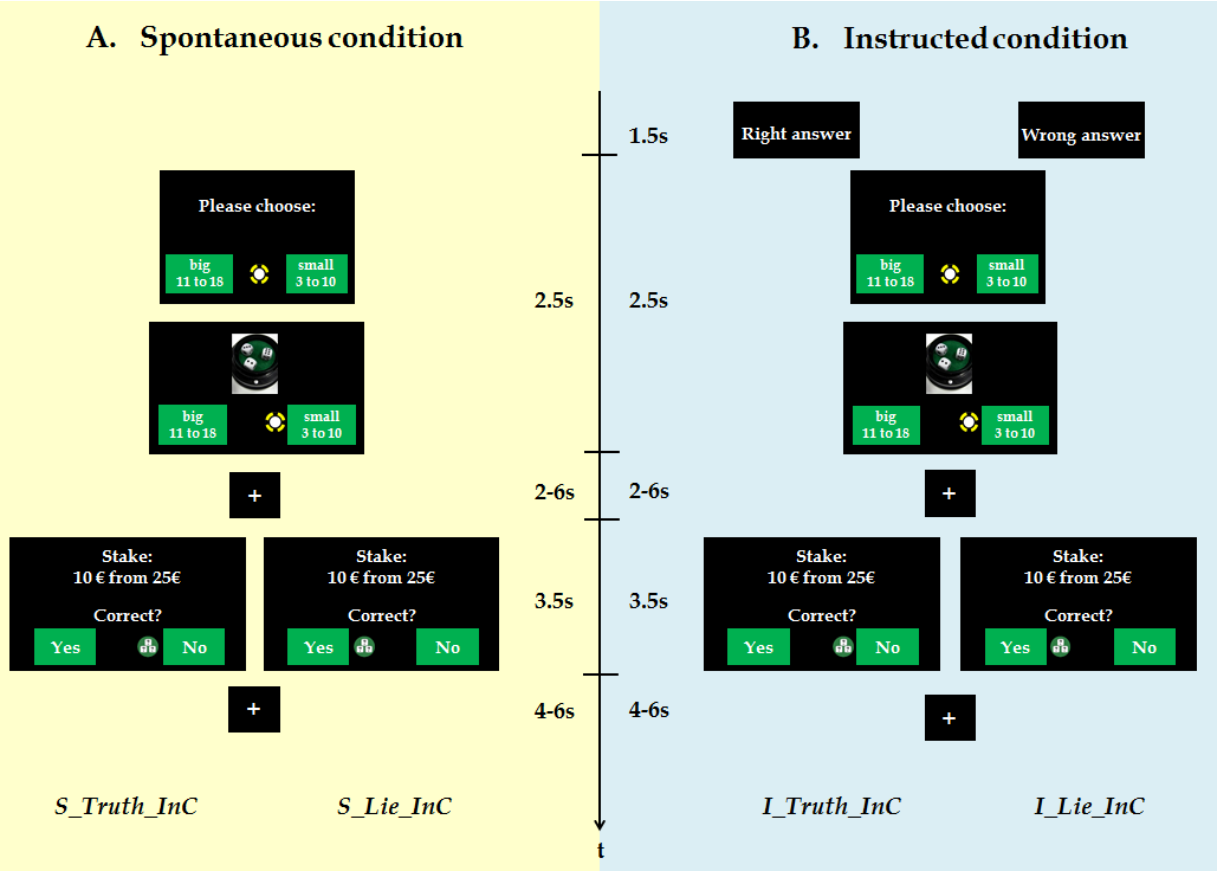

Figure 5.1: The experimental paradigm in Study 1. In the spontaneous session (A; marked in light yellow), a participant should first predict the result of the dice roll and bet on either "big" or "small" within 2.5s. The participant then freely reported his betting result within 3.5s. In this example, the result of the dice roll was "big," but the participant's prediction was "small." Thus, his prediction was wrong. In the instructed session (B; marked in light blue), the participant would first see the instruction (i.e., "Right answer" or "Wrong answer"). When the instruction was "Right answer," the participant should report his betting result truthfully (i.e., choosing "No"). When the instruction was "Wrong answer," the participant should report his betting result untruthfully (i.e., choosing "Yes"). In both sessions, if the participant's prediction was incorrect, choosing "No" would be "truth-telling" and choosing "Yes" would be "lying." (*S_Truth_InC*: spontaneous truth-telling in the trials with incorrect predictions; *S_Lie_InC*: spontaneous lying in the trials with incorrect predictions; *I_Truth_InC*: instructed truth-telling in the trials with incorrect predictions; *I_Lie_InC*: instructed lying in the trials with incorrect predictions)

5.2.3. Design and stimuli

The first fMRI session (i.e., spontaneous session) contained 162 trials in three scanning runs. There were nine different stakes (i.e., 0.5€, 1€, 1.5€, 9.5€, 10€, 10.5€, 19.5€, 20€, and 20.5€). Each of the nine different stakes repeated 18 times. The second fMRI session (i.e., instructed session) contained 180 trials in four scanning runs. Each of the nine different stakes repeated 20 times. Different amounts of trials in the instructed paradigm were designed to ensure that there were sufficient trials in each experimental condition. Each run lasted about 13 minutes. The spontaneous session started ahead of the instructed session, to avoid the possibility that the instructions of the instructed session influence the decision-making process in the spontaneous session. In total, the experimental stimuli consist of 342 different dice pictures. Half of the pictures showed "big" results and the other half showed "small" results.

5.2.4. Procedure

Before scanning, participants read the instructions of the spontaneous session. Participants were informed that the betting results they reported, rather than the actual betting results, would determine their final payoffs. In addition, the experimenters knew that there were opportunities for them to win stakes by choosing "Yes," regardless of their actual predictions, and there was no punishment if they respond incorrectly. After reading the instructions, participants performed one testing session and one practice session on the computer. The testing session was adopted to test participants' calculation ability. It contained 20 rounds of the calculation task. In each round, participants would see a picture of three dice. Within 2.5s, they should report "big" if the sum was from 11 to 18 and "small" if the sum was from 3 to 10. Participants with accuracy rates higher than 75% were allowed to perform the fMRI experiment. In the practice session, they completed 18 simulated trials to get familiar with the experiment. After that, participants entered the scanner and performed the first experimental session (i.e., spontaneous session).

After the spontaneous session, participants got out of the scanner and took a break of 30 minutes. During the break, participants were asked whether they knew that there were opportunities for them to earn more money by reporting the betting results untruthfully in the trials with incorrect predictions. Participants then filled in the questionnaires (see below for more details). After that, they read the instructions of the instructed session and completed 18 simulated trials on the computer. Participants performed the second fMRI session (i.e., instructed session) if the accuracy in the practice session was higher than 75%. After the instructed session, participants completed the questionnaires (see below for more details).

After each experimental session, based on a 9-point scale (1 = strongly disagree, 5 = neutral, 9 = strongly agree), participants gave their ratings to the following question: "How much do you agree with the following sentence: when the prediction was wrong, choosing 'Yes' in the experiment is a 'lie'." Participants also reported the emotional valence of different decisions (i.e., spontaneous lying and truth-telling, and instructed lying and truth-telling in the trials with incorrect predictions), based on Lang's Self-Assessment Manikin Valence Scale (Lang, 1980). The nine-level scale (1 = very unhappy, 5 = neutral, 9 = very happy) was adapted from PXLab (Irtel, 2008).

After the experiment, one trial from the spontaneous session and one trial from the instructed session were randomly chosen. Participants were paid accordingly. During the whole experiment, the words "cheat," "dishonest," "honest," "lie," and "truth" were not used in all of the instructions.

5.2.5. Data acquisition

Participants' responses in the scanner were collected via an MRI-compatible response device (NordicNeuroLab, Bergen, Norway). All images were collected on a Siemens Trio 3.0 Tesla scanner with a twelve-channel head coil. Structural scans included T1-weighted images (TR = 1570 ms; TE = 3.42 ms; flip angle = 15°; slice thickness =1.0 mm). The functional scans were collected using T2*-weighted echo planar images (EPI) pulse sequence employing a BOLD contrast (flip angle = 90°; TR = 2500 ms; TE

= 30 ms; field of view =192 × 192 mm²; 64 × 64 acquisition matrix; 37 slices with 3 mm slice thickness; in-plan resolution = 3 × 3 mm²).

The MRI scanner was upgraded to a Tim Trio System, after collecting data from 33 participants. Nine participants were scanned in the upgraded scanner.[19] Scans included T1-weighted images (TR = 1660 ms; TE = 2.75 ms; flip angle = 9°; slice thickness = 0.8 mm) and T2*-weighted echo planar images (flip angle = 90°; TR = 2500 ms; TE = 30 ms; field of view =192 × 192 mm²; 96 × 96 acquisition matrix; 37 slices with 3 mm slice thickness; in-plan resolution = 2 × 2 mm²).

5.2.6. Data analyses

Data from twelve participants were excluded: nine for technical reasons (image artifacts or excessive head movement of >3 mm or 3° of rotation); one for lack of attention in the experiment; one for not following the instructions in the instructed session; and one for not knowing that there were opportunities for him to earn more money by reporting incorrectly in the spontaneous session.

5.2.6.1. Behavioral data analyses

Statistical analyses of frequencies and reaction times for different decisions were conducted with SPSS 22.0 (IBM Corporation, Armonk, NY, USA). One sample t-tests, an independent sample t-test, and 2-by-2 repeated-measure analysis of variance (ANOVA) models were performed as indicated. All P values were two-tailed, and $P < 0.05$ was considered statistically significant. Post hoc analysis with Bonferroni correction was applied for significant interaction effects if any.

Based on the number of (un)truthful responses in the spontaneous trials with incorrect predictions and the consideration for sufficient trials (> 15) for fMRI data

[19] Among these nine participants, three participants behaved more honestly (honest group) and four participants behaved more dishonestly (dishonest group) in the spontaneous trials with incorrect predictions. Two participants were partially dishonest (partially dishonest group). Please see Session 5.2.6.1 and Session 5.3.1.1 for more details about the group classification.

analyses, participants were further classified into three groups: partially dishonest group, honest group, and dishonest group. Participants were assigned to the partially dishonest group, if they had at least 15 spontaneous lying trials with incorrect predictions and 15 spontaneous truth-telling trials with incorrect predictions. Participants were assigned to the honest group, if they had less than 15 spontaneous lying trials with incorrect predictions and more than 15 spontaneous truth-telling trials with incorrect predictions. Participants were assigned to the dishonest group, if they had less than 15 spontaneous truth-telling trials with incorrect predictions and more than 15 spontaneous lying trials with incorrect predictions.

5.2.6.2. Functional MRI data analyses

SPM8 was adopted for fMRI data analyses (Welcome Department of Cognitive Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm/). For each subject, EPI images were first realigned and resliced. Data sets that exhibited an overall movement of >3 mm or 3° of rotation in each run were not included. The anatomical image was co-registered with the mean EPI image of each participant which was further segmented. To create a template and normalize functional and anatomical scans to the MNI template, the SPM8's DARTEL tool was used. The normalized functional images were subsampled to 1.5 × 1.5 × 1.5 mm^3 and spatially smoothed using an 8-mm FWHM (full-width half maximum) Gaussian filter. High-pass temporal filtering with a cut-off of 128s was performed to remove low-frequency drifts.

Statistical analyses of the fMRI data were estimated using a general linear model (GLM). To investigate the neural correlates of (un)truthful responses in spontaneous and instructed paradigms, trials were categorized as the following conditions: (1) spontaneous lying in the trials with incorrect predictions (*S_Lie_InC*), (2) spontaneous truth-telling in the trials with incorrect predictions (*S_Truth_InC*), (3) spontaneous truth-telling in the trials with correct predictions (*S_Truth_C*), (4) instructed lying in the trials with incorrect predictions (*I_Lie_InC*), (5) instructed truth-telling in the trials with incorrect predictions (*I_Truth_InC*), and (6) instructed

truth-telling in the trials with correct predictions (*I_Truth_C*). Six regressors of interest above were included in the GLM, which contained the onsets of the reporting phase. Data from 19 participants who had sufficient trials in each condition (the partially dishonest group) were used. A parametric modulator of the betting values (i.e., stakes) for each regressor of interest was adopted. The onsets of the other events (i.e., trials with no response, the betting phase, spontaneous lying and instructed lying trials with correct predictions, and instructed trials with wrong responses) were combined into one *other* regressor. To remove the effects of head motion, six estimated head movement parameters were included. For the group-level analysis, four contrasts (i.e., *S_Lie_InC*, *S_Truth_InC*, *I_Lie_InC*, and *I_Truth_InC*) were entered into a flexible factorial model with two within-group factors (paradigm (the spontaneous paradigm and instructed paradigm) and decision (*Lie_InC* and *Truth_InC*)).

To further explore the neural processes involved in spontaneous truth-telling among participants with different levels of honesty, a similar GLM was built for those participants who were more honest (N = 15; the honest group). Considering the limited spontaneous lying trials with incorrect predictions (i.e., *S_Lie_InC*), five regressors of interest were included: (1) *S_Truth_InC*, (2) *S_Truth_C*, (3) *I_Lie_InC*, (4) *I_Truth_InC*, and (5) *I_Truth_C*. The onsets of the other events (i.e., trials with no response, the betting phase, spontaneous lying and instructed lying trials with correct predictions, instructed trials with wrong responses, and limited trials of S_*Lie_InC*) were combined into one *other* regressor. For the group-level analysis, a two-sample t-test was performed to compare the dishonest group (N = 19) with the honest group (N = 15) when they were making decisions to tell the truth in the spontaneous trials with incorrect predictions (*S_Truth_InC*).[20]

[20] The sample size in the dishonest group is not sufficient (N = 8). Therefore, the data from the dishonest group were not analyzed here.

A similar GLM was built for participants who behaved more dishonestly (N = 8; the dishonest group). Considering the limited spontaneous truth-telling trials with incorrect predictions (i.e., *S_Truth_InC*), five regressors of interest were included: (1) *S_Lie_InC*, (2) *S_Truth_C*, (3) *I_Lie_InC*, (4) *I_Truth_InC*, and (5) *I_Truth_C*. The onsets of the other events (i.e., trials with no response, the betting phase, spontaneous lying and instructed lying trials with correct predictions, instructed trials with wrong responses, and limited trials of *S_Truth_InC*) were combined into one *other* regressor. To check that if the instructed paradigm used in Study 1 is comparable to the instructed paradigms used previously, paired t-tests were performed to compare *I_Lie_InC* with *I_Truth_InC* and *I_Truth_C*. Since in previous instructed studies participants were not classified into different groups based on honesty levels, data from all 42 participants were used.

All results were voxel-level height uncorrected thresholded at $P < 0.001$ with spatial extent threshold set at $k = 50$, to take both type I and type II errors into account (Lieberman and Cunningham, 2009).

5.3. Results

5.3.1. Behavioral results

5.3.1.1. Lying frequencies in the spontaneous session

After the spontaneous session, all of the 42 participants whose data were used in fMRI data analyses, reported that they knew there was a possibility for them to win more money by choosing “Yes” when their predictions were incorrect.

In the spontaneous session, when participants correctly predicted the sum of the dice, they chose to respond honestly most of the time (N = 42; mean frequency of lying ± s.d.: 2.1% ± 2.83). In the 19 partially dishonest participants, the mean frequency of lying is 59.1% (s.d. = 16.5; ranged from 21% to 77%). In the remaining participants,

fifteen participants were more honest (mean frequency of lying ± s.d.: 3.1% ± 3.7; ranged from 0% to 13%), and eight participants were more dishonest (mean frequency of lying ± s.d.: 99.4% ± 0.5; ranged from 99% to 100%). The honesty levels here only describe participants' behaviors in the experiment. No conclusions could be drawn concerning their personalities or behavioral tendencies in general. In the 19 partially dishonest participants, the mean frequencies (± the standard deviation) of dishonest responses in the three stake ranges (0.5€ to 1.5€, 9.5€ to 10.5€, and 19.5€ to 20.5€) were 7.0% (± 8.0), 75.9% (± 31.0), and 91.8% (± 17.4) (Figure 5.2; red bars).

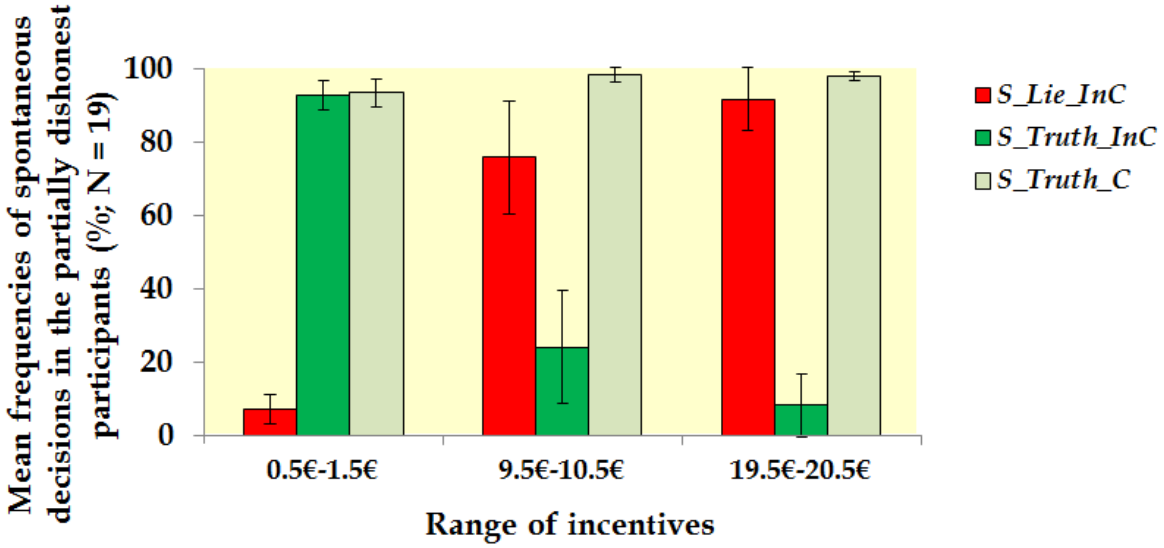

Figure 5.2: The behavioral results in Study 1. In 19 partially dishonest participants, the frequencies of lying and truth-telling in three betting value ranges are revealed. (*S_Lie_InC*: spontaneous lying in the trials with incorrect predictions; *S_Truth_InC*: spontaneous truth-telling in the trials with incorrect predictions; *S_Truth_C*: spontaneous truth-telling in the trials with correct predictions; error bars: s.d.)

5.3.1.2. Response accuracies in the instructed session

In the instructed session, the means and the standard deviations of response accuracies in all participants (N = 42) are listed in Table 5.1. To check whether participants paid enough attention to the task in the instructed session, the response accuracies in the four conditions (i.e., *I_Lie_InC*, *I_Lie_C*, *I_Truth_InC*, and *I_Truth_C*) were compared to the chance level (i.e., 50%). The accuracies in all four conditions were significantly higher than the chance level (all $ts_{(41)} > 21$, all ps < 0.001).

Table 5.1
Response accuracies (%; N = 42) in the instructed session

Conditions	**Mean (s.d.)**	
	Lying	**Truth-telling**
Incorrect predictions	91.3 (10)	93.5 (6)
Correct predictions	96 (5)	87.9 (11)

5.3.1.3. Judgments and emotional valence of (un)truthful responses

After each session, participants answered the question: "How much do you agree with the following sentence: when the prediction was wrong, choosing 'Yes' in the experiment is a 'lie'" based on a 9-point scale (1 = strongly disagree, 5 = neutral, 9 = strongly agree). The ratings of 42 participants in the spontaneous session (mean ± s.d: 6.98 ± 2.5) were significantly higher ($t_{(41)} = 4.88$, P < 0.001) than ratings in the instructed session (mean ± s.d: 4.45 ± 2.73; Figure 5.3). In the partially dishonest participants (N = 19), ratings in the spontaneous session (mean ± s.d: 7.05 ± 1.93) were also significantly higher than ratings in the instructed session (mean ± s.d: 4.42 ± 2.78; $t_{(18)} = 4.32$, P < 0.001).

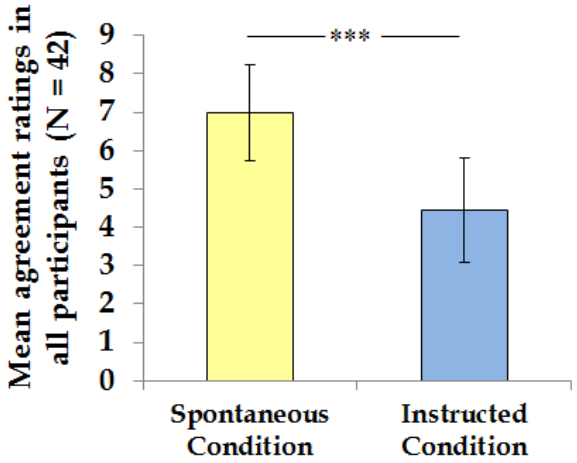

Figure 5.3: The behavioral results in Study 1. Participants' ratings toward the expression: "when the prediction was wrong, choosing 'Yes' in the experiment is a 'lie'" (1 = strongly disagree, 5 = neutral, 9 = strongly agree; ***P < 0.001; N = 42; error bars: s.d.).

To investigate whether there were differences in the emotional valences of lying and truth-telling between the spontaneous and instructed sessions, a 2 (decision: lying in the trials with incorrect predictions and truth-telling in the trials with incorrect predictions) × 2 (session: spontaneous and instructed sessions) ANOVA in 19 partially dishonest participants were performed.[21] The main effects of decision and session were not significant ($F_{(1, 18)}$ = 2.39, P = 0.14; $F_{(1, 18)}$ < 0.01, P > 0.99). A significant interaction effect of session × decision was observed ($F_{(1, 18)}$ = 13.10, P = 0.002; Figure 5.4). Post hoc analysis showed that the emotional valence of instructed lying (mean ± s.d: 6.4 ± 2.0) was higher than the emotional valence of instructed truth-telling (mean ± s.d: 5.0 ± 1.6) in the trials with incorrect predictions ($t_{(18)}$ = 2.79, P = 0.01; *I_Lie_InC* > *I_Truth_InC*), whereas the valence of spontaneous truth-telling (mean ± s.d.: 5.9 ± 1.9) was higher than the valence of spontaneous lying (mean ± s.d: 4.5 ± 1.4)

[21] Because of limited *S_Lie_InC* trials in honest participants and limited *S_Truth_InC* trials in dishonest participants, the analyses of valence were only performed in the participants from the partially dishonest group.

in the trials with incorrect predictions ($t_{(18)}$ = 2.17, P = 0.04; *S_Truth_InC* > *S_Lie_InC*).

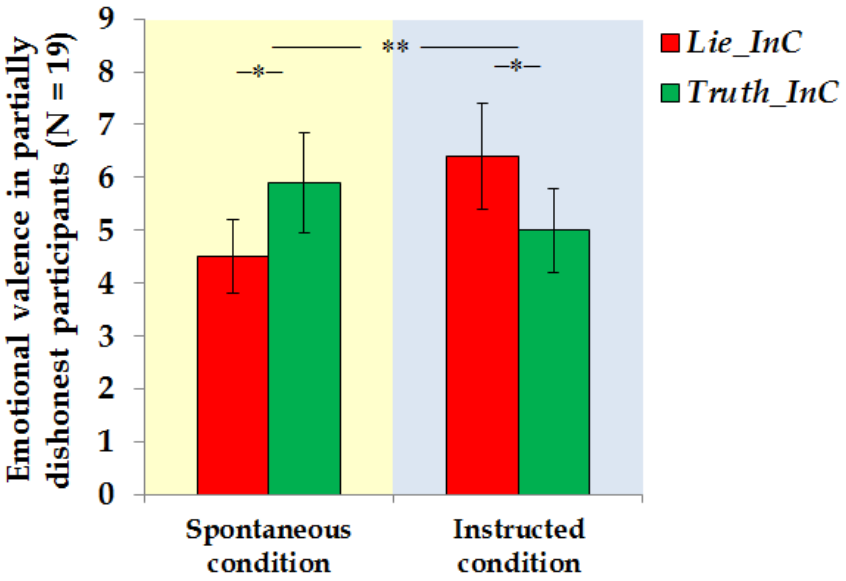

Figure 5.4: The behavioral results in Study 1. The emotional valences of lying and truth-telling in the spontaneous and the instructed trials with incorrect predictions (1 = very unhappy, 5 = neutral, 9 = very happy; *P < 0.05, **P < 0.01; N = 19; error bars: s.d.).

To investigate if there were differences in the reaction times for lying and truth-telling between two sessions, a 2 (decision: lying in the trials with incorrect predictions and truth-telling in the trials with incorrect predictions) × 2 (session: spontaneous and instructed sessions) ANOVA was conducted in the 19 partially dishonest participants.[22] A significant main effect of session was observed (spontaneous session > instructed session; $F_{(1, 18)}$ = 42.77, P < 0.001). The main effect of decision and the interaction effects were not significant ($F_{(1, 18)}$ = 3.31, P = 0.09; $F_{(1, 18)}$ = 0.19, P = 0.89). The means and the standard deviations of reaction times are listed in Table 5.2.

[22] Because of limited *S_Lie_InC* trials in the honest participants and limited *S_Truth_InC* trials in dishonest participants, the analyses of reaction times were only performed in the participants from the partially dishonest group.

Table 5.2
Reaction times in Study 1 (ms)

Conditions	Mean (s.d.)	
	Lie_InC	**Truth_InC**
Spontaneous session	1175 (275)	1226 (206)
Instructed session	963 (212)	1020 (247)

Lie_InC: lying in the trials with incorrect predictions.
Truth_InC: truth-telling in the trials with incorrect predictions.

5.3.2. Functional MRI results

5.3.2.1. Comparisons between instructed lying and instructed truth-telling

The contrasts of *I_Lie_InC* versus *I_Truth_InC* and *I_Lie_InC* versus *I_Truth_C* activated multiple regions (Table 5.3; N = 42). Most of the regions were also identified in previous studies that used instructed paradigms (Christ *et al.*, 2009), including the inferior frontal gyrus, the middle frontal gyrus, the anterior cingulate cortex, the insula, the inferior parietal lobule, and the supramarginal gyrus.

5.3.2.2. Interaction effects of decision and session

The results of a 2 (decision: *Truth_InC* and *Lie_InC*) by 2 (session: spontaneous and instructed sessions) whole-brain ANOVA analysis in 19 partially dishonest participants are listed in Table 5.4. No significant interaction effect of (*S_Lie_InC* - *S_Truth_InC*) versus (*I_Lie_InC* - *I_Truth_InC*) was observed. The opposite interaction effect of (*S_Truth_InC* - *S_Lie_InC*) versus (*I_Truth_InC* - *I_Lie_InC*) significantly activated the left precentral gyrus, the right ventral lateral prefrontal cortex (VLPFC), the dorsal lateral prefrontal cortex (DLPFC), the inferior parietal lobe (IPL), and the medial prefrontal gyrus (Figure 5.5A). Post-hoc analyses showed that activity in the right VLPFC, the right DLPFC, and the right IPL was significantly higher in the *S_Truth_InC* condition than activity in the *S_Lie_InC* condition. No significant difference was observed in the contrast of *I_Lie_InC* versus *I_Truth_InC* (Figure 5.5B).

Table 5.3
Brain activation in the comparison between instructed lying and instructed truth-telling in all 42 participants

Condition	Hem	Voxel	Brain area	MNI coordinates			Z value
				x	y	z	
I_Lie_InC > *I_Truth_InC*							
	L	66	Middle frontal gyrus	-41	18	50	3.81
	L	304	Anterior cingulate gyrus	-5	37	23	3.76
	L	122	Superior frontal gyrus	-5	9	69	3.6
	L	104	Superior frontal gyrus	-9	30	56	3.49
	L	251	Supramarginal gyrus	-54	-57	35	3.47
	R	63	Sub-gyral	20	12	38	3.43
I_Lie_InC < *I_Truth_InC*							
	R	201	Middle occipital gyrus	45	-74	26	4.29
I_Lie_InC > *I_Truth_C*							
	L	1109	Middle frontal gyrus	-38	18	41	4.85*
	L	277	Inferior frontal gyrus	-48	24	-4	4.55
	L	1233	Inferior parietal lobule	-47	-56	47	4.24*
	R	86	Inferior frontal gyrus	50	24	-6	4.22
	L	469	Sub-gyral	-15	22	42	3.84
	L	309	Supplementary motor area	-2	27	56	3.76
	L	276	Superior frontal gyrus	-20	42	36	3.75
	L	148	Superior frontal gyrus	-26	54	20	3.69
	R	152	Middle frontal gyrus	29	50	17	3.67
	R	52	Middle frontal gyrus	35	53	5	3.56
I_Lie_InC < *I_Truth_C*							
	R	679	Middle temporal gyrus	47	-74	26	4.38*
	M	250	Anterior cingulate cortex	0	53	-1	3.83

Results were all voxel-level height thresholded at $P < 0.001$, $k > 50$ voxels, uncorrected.

* survived after cluster-level family wise error (FWE) correction, $P_{FWE\text{-}corrected} < 0.05$.

I_Lie_InC: instructed lying in the trials with incorrect predictions.

I_Truth_InC: instructed truth-telling in the trials with incorrect predictions.

I_Truth_C: instructed truth-telling in the trials with correct predictions.

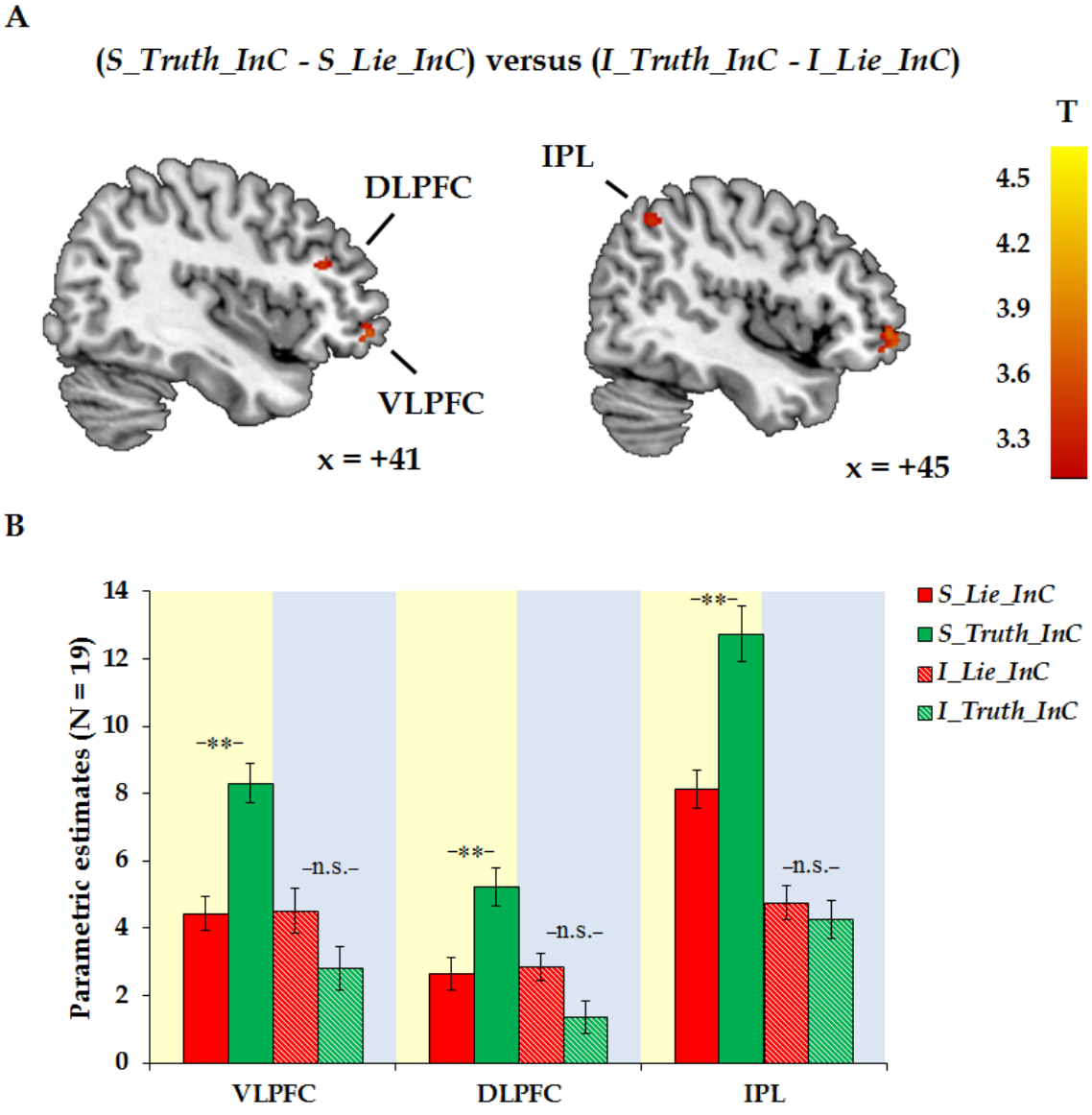

Figure 5.5: fMRI results in Study 1. (A) In the partially dishonest participants (N = 19), the right VLPFC, the right DLPFC, and the right IPL were significantly activated in the contrast of (*S_Truth_InC - S_Lie_InC*) versus (*I_Truth_InC - I_Lie_InC*) ($P < 0.001$, $k > 50$, uncorrected). (B) Parametric estimates were extracted from the whole cluster in the three regions. (VLPFC: ventral lateral prefrontal cortex; DLPFC: dorsal lateral prefrontal cortex; IPL: inferior parietal lobule; **: $P < 0.01$, n.s.: not significant; error bars: s.e.m.)

Table 5.4
fMRI results of a whole brain ANOVA analysis in 19 partially dishonest participants

L/R	Voxel	Brain area	MNI coordinates			Z value
			x	y	z	
Main effect: spontaneous session > instructed session						
R	1444	Inferior frontal gyrus/anterior insula	47	21	-3	4.58*
L	1973	Inferior frontal gyrus/anterior insula	-38	23	-7	4.55*
R	2882	Inferior parietal lobule	47	-44	54	4.52*
R	1585	Middle frontal gyrus	35	51	21	4.31*
R	959	Anterior cingulate cortex	3	30	30	4.19*
L	1666	Inferior parietal lobule	-38	-48	42	4.10*
R	1009	Medial globus pallidus	17	-6	2	4.08*
L	833	Middle frontal gyrus	-47	37	23	4.06
R	461	Middle occipital gyrus	33	-86	3	3.90
L	189	Post cingulate cortex	-2	-21	30	3.80
L	615	Precuneus	-5	-68	59	3.80
L	88	Medial globus pallidus	-14	-3	0	3.70
R	308	Superior frontal gyrus	6	21	65	3.61
L	52	Middle occipital gyrus	-27	-89	-3	3.38
Main effect: instructed session > spontaneous session						
R	132	Parietal lobe	60	-17	47	4.36
Main effect: Lie_InC > Truth_InC						
L	463	Subgenual anterior cingulate cortex	-6	30	-3	4.28
R	58	Brainstem	5	-36	-23	3.41
L	61	Medial frontal gyrus	-6	55	24	3.41
L	90	Angular	-44	-72	36	3.35
Main effect: Truth_InC > Lie_InC						
R	221	Middle occipital gyrus	42	-45	56	3.54
Interaction: (S_Lie_InC - S_Truth_InC) - (I_Lie_InC - I_Truth_InC)						
None						
Interaction: (S_Truth_InC - S_Lie_InC) - (I_Truth_InC - I_Lie_InC)						
L	84	Precentral gyrus	-9	-27	72	3.97
R	150	Ventral lateral prefrontal cortex	43	51	-9	3.64
R	77	Inferior parietal lobule	48	-41	42	3.60
R	106	Medial frontal gyrus	15	43	30	3.50
R	77	Inferior parietal lobule	48	-56	48	3.44
R	97	Dorsal lateral prefrontal cortex	39	31	23	3.43

Results were voxel-level height thresholded at $P < 0.001$, $k > 50$ voxels, uncorrected.

* survived after cluster-level family wise error (FWE) correction, $P_{FWE\text{-}corrected} < 0.05$.

S: the spontaneous session/paradigm; *I*: the instructed session/paradigm.

InC: the trials with incorrect predictions; *C*: the trials with correct predictions.

5.3.2.3. Comparisons between spontaneous truth-telling in partially dishonest participants and spontaneous truth-telling in honest participants

To explore the neural correlates of spontaneous truth-telling in participants with different levels of honesty, partially dishonest participants' (N = 19) decisions of spontaneous truth-telling with incorrect predictions (i.e., *S_Truth_InC*) were compared to the decisions of honest participants (N = 15). Several brain regions, particularly the right VLPFC, the right DLPFC, and the right IPL, showed increased BOLD signals in the partially dishonest participants, compared with honest participants. None survived in the opposite contrast (Table 5.5).

Table 5.5
fMRI results of the comparisons between spontaneous truth-telling in partially dishonest participants and spontaneous truth-telling in honest participants

L/R	Voxel	Brain area	MNI coordinates			Z value
			x	y	z	
S_Truth_InC: partially dishonest group > honest group						
L	392	Inferior frontal gyrus/anterior cingulate cortex	-35	23	-11	4.46
R	580	Superior frontal gyrus	20	39	45	4.21*
L	469	Inferior parietal lobule	-51	-57	45	4.14
L	430	Middle frontal gyrus	-39	30	39	3.95
R	94	Ventrolateral prefrontal cortex	42	48	-11	3.72
R	104	Dorsolateral prefrontal cortex	38	45	18	3.61
R	58	Superior parietal lobule	21	-65	51	3.56
R	53	Superior parietal lobule	36	-48	60	3.52
L	53	Middle frontal gyrus	-42	51	5	3.47
R	55	Inferior parietal lobule	50	-54	45	3.39
S_Truth_InC: partially dishonest group < honest group						
None						

Results were voxel-level height thresholded at $P < 0.001$, $k > 50$ voxels, uncorrected.
* survived after cluster-level family wise error (FWE) correction, $P_{FWE\text{-}corrected} < 0.05$.
S_Truth_InC: spontaneous truth-telling in the trials with incorrect predictions.

5.4. Discussion

Study 1 is the first fMRI study to directly compare the (un)truthful responses in the spontaneous paradigm with the (un)truthful responses in the instructed paradigm. The essential difference between the spontaneous paradigm and the instructed paradigm is whether individuals can freely make their decisions to lie. Based on participants' judgments of untruthful responses, spontaneous untruthful responses were thought to be closer to the concept of "lies" compared with instructed untruthful responses. The results suggest that untruthful responses in the instructed paradigm differ from untruthful responses in the spontaneous paradigm. The notion was further supported by the different emotional valences in two paradigms. In the spontaneous session, the emotional valence that occurred when participants were telling the truth was significantly more positive than when they were lying, whereas the opposite trend was observed in the instructed session, namely, lying was more positive than truth-telling. The more negative valence observed during spontaneous lying compared with truth-telling might be due to high psychological costs of lying (e.g., positive self-image damage, guilt or aversion) (Gneezy, 2005; Mazar and Ariely, 2006; Mazar *et al.*, 2008; Ellingsen *et al.*, 2010; Shalvi *et al.*, 2010; Battigalli *et al.*, 2013; Gneezy *et al.*, 2013; Ploner and Regner, 2013). The reversed patterns observed in the instructed session might be due to great monetary gains and low psychological costs of lying. These behavioral findings suggest that the ecological validity and the emotional responses to lying and truth-telling differ in these two paradigms.

At the neural level, if (un)truthful responses in the two paradigms share the similar mental process and neural mechanisms, no significant interaction effect of the paradigm and decision would be expected. However, the effect of the interaction (i.e., (*S_Truth_InC-S_Lie_InC*) versus (*I_Truth_InC-I_Lie_InC*)) significantly activated the right DLPFC, the right VLPFC, and the right IPL. These regions were significantly activated in the comparison of spontaneous truth-telling versus spontaneous lying, but not in the comparison of instructed truth-telling versus instructed lying. Moreover,

compared with the honest participants, spontaneous truth-telling elicited higher activity in the right DLPFC, the right VLPFC, and the right IPL in dishonest participants. These three regions belong to a frontoparietal control system that is associated with cognitive control in decision-making processes (Vincent *et al.*, 2008). The DLPFC has been reported to be involved in tasks requiring cognitive control and the inhibition of prepotent impulses (Sanfey *et al.*, 2003; Aron *et al.*, 2004; Spitzer *et al.*, 2007). Our results are consistent with the previous findings in the study of Greene and Paxton (2009), suggesting that honest individuals required less cognitive effort to make honest decisions than dishonest individuals (Greene and Paxton, 2009; Abe and Greene, 2014).

When using instructed paradigms, the complex executive functions associated with (dis)honest decision-making might not be fully investigated (Sip *et al.*, 2008). The findings from Study 1 suggest two main points: 1) the neural processes involved in (dis)honest decision-making in the instructed paradigm differ from the neural processes involved in the spontaneous paradigm, particularly in the frontoparietal network and 2) in the incentivized and more natural setting, when individuals choose to tell the truth, extra cognitive control and greater involvement of the associated brain regions might be required to resist dishonest gain.

5.5. Limitations

There are some limitations in Study 1. First, although the experimental design in the spontaneous session allowed participants to freely make their decisions, they knew that their responses were being observed by others. The decision-making process might be different from the process in natural settings in which they can conceal their

lies[23]. Second, the instructed session was always conducted after the spontaneous session. The fixed order might induce habituation in the instructed session. A comparison between instructed lying and truth-telling was performed to investigate whether the instructed session could provide similar results as the previous instructed studies. Most of the activated brain regions were also identified in the previous instructed studies. Therefore, the instructed paradigm used in Study 1 is comparable to the previous instructed paradigms. Finally, the absence of other important factors, such as social interaction (Lisofsky *et al.*, 2014), still induces substantial gaps between lies in real life and lies in the spontaneous paradigm in Study 1.

[23] For example, in the study by Greene and Paxton (2009), participants' predictions about the coin flips were not recorded in the condition with cheating opportunity. Therefore, dishonest participants could conceal their lying behaviors. But their study also suffers the drawback that researchers could not identify lying and truth-telling trials.

6. Study 2: The neural mechanisms underlying the modulation of altruistic outcomes on communicators' deceptive decision-making process

Collaborators on the research described in this section are Yang Hu, Dennis Dynowski, Jian Li, and Bernd Weber. (The running title of the manuscript: "The good lie: altruistic goals modulate processing of deception in the anterior insula." Manuscript under review.)

Author contributions: L.Y., Y.H., J.L., and B.W. designed research; L.Y., Y.H., and D.D. performed research; L.Y., Y.H., and J.L. analyzed data.

6.1. Introduction

Empirical evidence supports the hypothesis that there are psychological costs of lying. Introducing honesty concerns in the economic decision-making process (i.e., lying could lead to higher monetary gains) decreases people's tendency to pursue their own interests (Gneezy, 2005; Shalvi *et al.*, 2011b). Such aversion can be further modulated by altruistic outcomes. For example, lying behavior increased if it benefited charities (Lewis *et al.*, 2012). Altruistic outcomes might promote deception by overcoming the psychological barrier of lying.

However, the neural mechanisms underlying the impact of altruistic outcomes on lying remain unknown. The aim of Study 2 is to investigate this research question. In Study 1, the behavioral and neural findings suggest that the spontaneous paradigms might be more suitable to investigate deception compared with the instructed paradigms. Therefore, Study 2 used the task that allows participants to freely make their decisions. An interactive game was adopted to better investigate the altruistic impact on deception and to overcome the limitations in Study 1 (particularly the lack

of social interaction). The sender-receiver game (Gneezy, 2005) was briefly introduced in Section 1.3. In the original version of the game, a sender has private information of two possible monetary distributions. The sender transmits the message about the advantageous option for the receiver. The receiver decides to implement one of the two options, purely based on the message. The game creates an incentivized context in which the sender can send an untruthful or truthful message to achieve different payoff outcomes. However, the potential problem is that truth-telling behaviors might be "sophisticated deception," meaning that a sender could send a truthful message with the expectation that the receiver does not believe the message and chooses the allocation that is advantageous to the sender (Sutter, 2009). In Study 2, the message conveyed from the sender is about a computer's random choice of the payoff option and does not contain any information about the advantageous option. The receiver does not directly choose to implement one of the two payoff options. Instead, the receiver chooses whether to believe the message. If the receiver believes the message, the payoff option chosen by the sender would be implemented. If the receiver does not believe the message, both players earn nothing. Two conditions were introduced, with either charities or senders themselves as the beneficiaries (i.e., charity-profit condition and self-profit condition) to test the effect of other- and self-profit outcomes. In the condition with honesty concerns to achieve higher payoffs, the senders could benefit themselves or a charity more by deceiving the receiver. In the study by Gneezy (2005), he compared the rates of choosing self-benefiting allocations in a dictator game to the rates in the sender-receiver game to rule out the preferences for the payoff distribution and estimate the extent of honesty concerns. In a similar vein, in Study 2, a control condition was performed, in which the senders could achieve higher payoffs via telling the truth (i.e., condition with no honesty concerns).

6.2. Materials and methods

6.2.1. Participants

Forty-seven healthy participants (29 females; mean ± s.d. age = 25.77 ± 3.71 years ranged from 19 to 35 years) took part in the fMRI experiment. All participants reported no prior history of psychiatric or neurological disorders and had normal or corrected-to-normal vision. All participants gave informed consent, and the study was approved by the Ethics Committee of the University of Bonn.

6.2.2. Tasks

Study 2 adopted a modified sender-receiver paradigm (Figure 6.1). Participants, as senders, played the game with an anonymous receiver. In each trial, participants would first see two payoff options A and B. Each payoff option consisted of the payoff for the sender, which was represented by the blue bar, and the payoff for the receiver, which was represented by the red bar. If an icon of a pre-selected charitable organization was displayed, the sender's payoff would be donated to the charity (i.e., the charity-profit condition; Figure 6.1A). If a blue silhouette was displayed, participants would be paid based on the sender's payoff (i.e., the self-profit condition; Figure 6.1B). One payoff option was randomly chosen by a computer (indexed by an icon of the computer). By pressing the button on the response grips within 4s, participants selected one of the payoff options to phrase a message: "The computer chose option x to be implemented" (where x = A or B). After pressing the button on the response grips, a yellow frame appeared to index the corresponding choice for 0.5s. A fixation cross was then displayed for a jittered interval (i.e., 8-10s minus the reaction time for that trial).

After scanning, the computer would randomly select one trial with participants as the beneficiary and one trial with a pre-selected charity as the beneficiary. The receiver would receive the messages in the two selected trials, and the receiver chose one of the two options: believe or not believe. If the receiver chose to believe the messages, the

option selected by the participant would be implemented (Figure 6.1C). Otherwise, both the sender and the receiver would earn nothing (Figure 6.1D). After the receiver made the decisions, information including the payoff options, whether the sender lied or not, and the final implemented options would be presented to the receiver. The yellow frame indexed the final implemented payoff option. Crucially, from the sender's perspective, if the computer chose the option with a high payoff for the senders, participants did not need to lie to get the higher payoff (i.e., no honesty concerns to get higher payoffs condition). Otherwise, participants would have honesty concerns to get the higher payoff (i.e., honesty concerns to get higher payoffs condition; Figure 6.1A and B).

6.2.3. Design and stimuli

The current event-related fMRI study adopted a 2 (beneficiary: participants or charitable organizations; abbreviation: *Self* or *Charity*) by 2 (honesty concerns: with or without honesty concerns to get higher payoffs; abbreviation: *HonCon* or *NoHonCon*) within-subject factorial design. One 40-min scanning run contained 192 trials in total (48 trials per condition). In each trial, the payoff for the sender in one option was lower than the other one (Table 6.1). The low payoff for the sender was drawn out of the three monetary amounts (2€, 6€, or 10€). The high payoff for the sender was built based on the low payoff (2€, 6€, or 10€) with one of the following increments: 1€, 4€, 6€, 8€, and 15€ (abbreviation of the payoff difference: *PD*). The combination with the *PD* of 4€, 6€, and 8€ repeated four times and others (i.e., the *PD* of 1€ and 15€) repeated twice. The display positions of the high and low payoff options were counterbalanced within each participant. In half of the trials, the payoff for the receiver in both options was equal to the high payoff for the sender. In the other half of the trials, the receiver's payoff was equal to the low payoff for the sender.

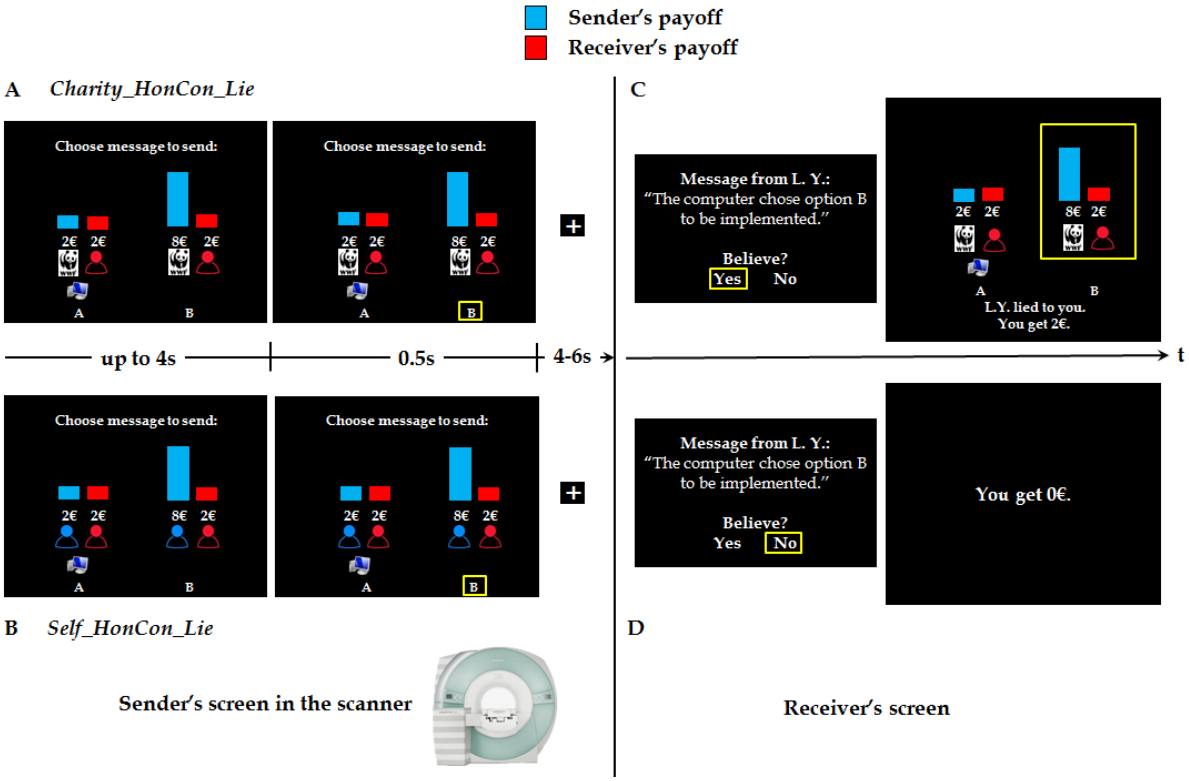

Figure 6.1: The experimental paradigm of the modified sender-receiver game in Study 2. A participant in the scanner played the game as a sender. The blue and the red bars represent the payoff for the sender and the payoff for an anonymous receiver, respectively. In this example, a computer chose the option with a low payoff for the sender (i.e., option A; indexed by the computer icon), meaning that the participant had honesty concerns to get the high payoff. The participant chose one of two payoff options to phrase a message (e.g., "The computer chose option B to be implemented"). The payoff for the sender would be donated to a pre-selected charity (A; indexed by the charity icon; *Charity_HonCon_Lie*) or obtained by the participant (B; indexed by the blue silhouette; *Self_HonCon_Lie*). After the scanning, an anonymous receiver would receive the messages from two randomly selected trials. If the receiver believed participant's message (C), the option chosen by the participant would be implemented (the option marked by the yellow frame). If the receiver did not believe (D), both the receiver and the sender would earn 0€. (*Charity_HonCon_Lie*: lying in the condition with honesty concerns to get higher payoffs and with a charity as the beneficiary; *Self_HonCon_Lie*: lying in the condition with honesty concerns to get higher payoffs and with participants as the beneficiary.)

Table 6.1
The payoff structure in Study 2

Low / High / PD	2	6	10
1	3	7	11
4	6*	10*	14*
6	8*	12*	16*
8	10*	14*	18*
15	17	21	25

Low: low payoffs for the sender; High: high payoffs for the sender.
PD: the payoff difference between the high payoff and the low payoff for the sender.
The combination marked by * repeated four times in each condition and others repeated twice in each condition.

6.2.4. Procedure

Before the fMRI experiment, all participants read a brief introduction about 6 charitable organizations (i.e., Doctors without Borders, World Wide Fund for Nature, Greenpeace, Amnesty International, German World Hunger Aid, and The United Nations Children's Emergency Fund) and selected one organization to be the beneficiary as was done previously (Kuss *et al.*, 2013). They read the instructions of the experiment and completed a manipulation check. Every participant signed a consent form to authorize the experimenters to donate the money (i.e., the sender's payoff in one randomly selected charity-profit trial) to the pre-selected charity. Participants were informed that they would play the game with an anonymous receiver, and neither of them would know the identity of the other. They would not meet the receiver during the whole experiment. Only the initials of participants would be presented to the receiver. Unknown to the participants, the receiver in the experiment was a confederate and chose to believe all messages.

In the scanner, to get familiar with the experiment, participants first completed a practice session, which included eight simulated trials. After the practice, participants entered the scanner and made a series of decisions to phrase messages. After the fMRI experiment, one trial from the charity-profit condition and one trial from the self-profit condition would be randomly chosen. Participants were finally informed of the total payoff (i.e., 10€ participation fee plus the payoff from the self-profit condition) and were paid accordingly. Money was donated to the corresponding charitable organizations according to participants' choices in the selected charity-profit trials.

6.2.5. Data acquisition

Participants' responses in the scanner were collected via an MRI-compatible response device (NordicNeuroLab, Bergen, Norway). All images were acquired on a Siemens Trio 3.0-Tesla scanner with a standard 32-channel head coil. Structural scans included T1-weighted images (TR = 1660 ms; TE = 2.75 ms; flip angle = 9°; slice thickness = 0.8 mm). The functional scans were acquired using T2*-weighted echo planar images (EPI) pulse sequence employing a BOLD contrast (flip angle = 90°; TR = 2500 ms; TE = 30 ms; 96 × 96 acquisition matrix; field of view =192 × 192 mm^2; 37 slices; in-plane resolution = 2 × 2 mm^2, thickness = 3 mm).

6.2.6. Data analyses

Data from five subjects were excluded, due to the following reasons: two for excessive head movements (i.e., > 3 mm or 3° of rotation), two for technical failure during scanning, and one due to a misunderstanding of the experiment. All following analyses were based on the data of remaining 42 participants (26 females; mean ± s.d. age = 25.45 ± 3.43 years ranged from 19 to 33 years).

6.2.6.1. Behavioral data analyses

Statistical analyses of percentages and reaction times were conducted with SPSS 22.0 (IBM Corporation, Armonk, NY, USA). Two-sided Wilcoxon signed-rank tests were performed. All P values were two-tailed, and $P < 0.05$ was considered statistically significant.

Measure of the impact of altruistic goals on honesty concerns

The effects of honesty concerns in the charity-profit and the self-profit conditions were estimated by the ratios of payoff loss caused by honesty concerns. The higher the effect of honesty concerns, the higher the ratio of payoff loss. To estimate the impact of altruistic outcomes on honesty concerns, the ratios of payoff loss in the self-profit and the charity-profit conditions would be compared. If altruistic outcomes reduce honesty concerns to a greater extent, the difference in the payoff loss ratios between the self-profit and the charity-profit condition would be larger.

In Gneezy's experiment (2005), the rates of choosing advantageous allocations in the deception game and the rates in the the dictator game were compared to estimate the extent of lying aversion by ruling out the preferences of the payoff distribution. In Study 2, the *NoHonCon* condition was introduced as the control condition where participants could get higher payoffs via truth-telling. In both the *HonCon* and the *NoHonCon* conditions, participants could earn more by choosing the high payoff option to phrase a message. In the *HonCon* condition, however, choosing the high payoff option corresponds to sending an untruthful message to deceive the receiver.

To some extent, participants had honesty concerns in the *HonCon* condition.[24] The individual who has higher honesty concerns tends to choose the high payoff option less often in the honesty concerns condition than the condition without honesty concerns, resulting in higher payoff loss. Therefore, the payoff loss in the *HonCon* condition was compared with that in the *NoHonCon* condition, and the ratios of the payoff loss caused by honesty concerns were calculated as follows:

$$HC_{PL:\,Charity} = \frac{\sum_{i=1}^{5}(N_{PD_i\_Charity\_NoHonCon\_Truth} - N_{PD_i\_Charity\_HonCon\_Lie}) \cdot PD_i}{\sum_{i=1}^{5} N_{PD_i\_Charity\_NoHonCon\_Truth} \cdot PD_i} \quad 6.1$$

$$HC_{PL:\,Self} = \frac{\sum_{i=1}^{5}(N_{PD_i\_Self\_NoHonCon\_Truth} - N_{PD_i\_Self\_HonCon\_Lie}) \cdot PD_i}{\sum_{i=1}^{5} N_{PD_i\_Self\_NoHonCon\_Truth} \cdot PD_i} \quad 6.2$$

In the equation (6.1), $N_{PD_i\_Charity\_NoHonCon\_Truth}$ denotes the numbers of truth-telling decisions in the trials with different payoff differences PD_i (i.e., 1, 4, 6, 8, or 15) and with a charity as the beneficiary in the situation without honesty concerns. The Parameter $N_{PD_i\_Charity\_HonCon\_Lie}$ denotes the numbers of lying decisions in the trials with different payoff differences PD_i (i.e., 1, 4, 6, 8, or 15) and with a charity as the beneficiary in the situation with honesty concerns. $\sum_{i=1}^{5} N_{PD_i\_Charity\_HonCon\_Lie} \cdot PD_i$ and $\sum_{i=1}^{5} N_{PD_i\_Charity\_NoHonCon\_Truth} \cdot PD_i$ denote the total payoff increments when participants chose high payoff options. Choosing high payoff options is lying in the *HonCon* condition and is truth-telling in the *NoHonCon* condition. The rules hold true for the equation (6.2) except that the beneficiaries were participants themselves (i.e., the self-profit condition). $HC_{PL:\,Charity}$ and $HC_{PL:\,Self}$ respectively denoted the

[24] In Session 6.1, 6.2.2, and 6.2.3, the *HonCon* condition refers to the situation with honesty concerns to earn high payoffs, where a computer chooses the low payoff option and participants can only earn the high payoff by lying. The honesty concerns condition might have different impacts on different individuals. For example, some individuals do not have concerns regarding honesty in the *HonCon* condition and always choose higher payoffs, regardless of the means (lying in the *HonCon* condition or truth-telling in the *NoHonCon* condition). Therefore, honesty concerns in the experimental design describe the situation in which higher payoffs can be obtained by lying. The effect of honesty concerns estimated in every participant describes the extent to which the *HonCon* situation influences participants' decisions.

ratios of the payoff loss (abbrev.: *PL*) caused by honesty concerns in the charity-profit and the self-profit conditions. Higher values of $HC_{PL:\ Charity}$ and $HC_{PL:\ Self}$ indicate larger ratios of payoff loss caused by honesty concerns in the charity-profit and self-profit conditions, respectively.

To estimate the impact of altruistic outcomes on the effect of honesty concerns, the difference between $HC_{PL:\ Self}$ and $HC_{PL:\ Charity}$ was calculated as follows:

$$HC_{PL:\ Self-Charity} = HC_{PL:Self} - HC_{PL:\ Charity} \qquad 6.3$$

$HC_{PL:\ Self\text{-}Charity}$ in the equation (6.3) denotes the impact of altruistic outcomes on the ratios of payoff loss caused by honesty concerns. If $HC_{PL:\ Self\text{-}Charity} = 0$, the ratios of payoff loss caused by honesty concerns are indifferent between the charity-profit and the self-profit conditions. If $HC_{PL:\ Self\text{-}Charity} < 0$, the charity-profit condition increases the ratios of payoff loss. If $HC_{PL:\ Self\text{-}Charity} > 0$, the charity-profit condition decreases the ratios of payoff loss.

If there was a participant who always chose high payoff options in both the *Charity_NoHonCon* condition and the *Charity_HonCon* condition, the equation (6.1) would be:

$$HC_{PL:\ Charity} = \frac{(6-6)\cdot 1+(12-12)\cdot 4+(12-12)\cdot 6+(12-12)\cdot 8+(6-6)\cdot 15}{6\cdot 1+12\cdot 4+12\cdot 6+12\cdot 8+6\cdot 15} = 0.$$

If the same participant chose high payoff options in the *Self_NoHonCon* condition regardless of different *PD*s, and chose high payoff options in 100% of the *Self_HonCon* trials with the *PD* of 15, 50% of the *Self_HonCon* trials with the *PD* of 4, 6, and 8, and 0% of the *Self_HonCon* trials with the *PD* of 1. The equation (6.2) would be:

$$HC_{PL:\ Self} = \frac{(6-0)\cdot 1+(12-6)\cdot 4+(12-6)\cdot 6+(12-6)\cdot 8+(6-6)\cdot 15}{6\cdot 1+12\cdot 4+12\cdot 6+12\cdot 8+6\cdot 15} \approx 0.37.$$

The impact of altruistic outcomes on the ratios of payoff loss caused by honesty concerns in this participant would be: $HC_{PL:\ Self-Charity} = 0.37 - 0 = 0.37$, meaning

that the charity-profit condition decreases the ratios of payoff loss caused by honesty concerns compared with the self-profit condition.

6.2.6.2. Functional MRI data analyses

SPM8 was used for fMRI data analyses (Welcome Department of Cognitive Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm/). For each subject, EPI images were first realigned and corrected for slice timing. Data sets that exhibited movement of >3 mm or 3° of rotation were not included. The anatomical image was co-registered to the mean EPI image. It was segmented, and parameters for normalization to MNI space were generated. EPI data were then projected onto MNI space with a 2×2×2 mm^3 resolution and smoothed using an 8-mm FWHM Gaussian filter. High-pass temporal filtering with a cut-off of 128s was performed to remove low-frequency drifts.

The statistical analyses of fMRI data were based on three general linear models (i.e., GLMs 1-3). Based on the events of interest, trials were categorized as the following conditions: *Charity_HonCon_Lie* (lying in the condition with honesty concerns to get higher payoffs and with a charity as the beneficiary; Figure 6.1A), *Self_HonCon_Lie* (lying in the condition with honesty concerns to get higher payoffs and with oneself as the beneficiary; Figure 6.1B), *Charity_NoHonCon_Truth* (truth-telling in the condition without honesty concerns to get higher payoffs and with a charity as the beneficiary), *Self_NoHonCon_Truth* (truth-telling in the condition without honesty concerns to get higher payoffs and with oneself as the beneficiary), *Charity_HonCon_Truth* (truth-telling in the condition with honesty concerns and with a charity as the beneficiary), and *Self_HonCon_Truth* (truth-telling in the condition with honesty concerns and with oneself as the beneficiary). Given that there are very limited trials in the two conditions *of NoHonCon_Lie* (i.e., lying in the condition without honesty concerns to get higher payoffs; mean trials in charity-profit condition: 3.24; mean trials in self-profit condition: 4.98), these two conditions were not used in the fMRI analysis.

GLM 1 was set up to investigate whether lying and truth-telling to get higher payoffs are encoded differently in the charity-profit and the self-profit conditions, as well as the neural correlates that reflect individual differences in the impact of altruistic goals on honesty concerns. Four regressors of interest were included, which contained the onsets of four events: (1) *Charity_HonCon_Lie*, (2) *Charity_NoHonCon_Truth*, (3) *Self_HonCon_Lie,* and (4) *Self_NoHonCon_Truth*. In addition, one parametric modulator of payoff differences was added for each regressor respectively to remove the confounding effects that might be driven by 1) advantageous and disadvantageous inequality situations (Hsu *et al.*, 2008; Yu *et al.*, 2014) and 2) the payoff differences between the low payoff and the high payoff for the sender. Onsets of the other events (i.e., trials with no response and trials with decisions that lead to low payoff outcomes to senders) were combined into one regressor of no interest (i.e., *other* regressor). Six estimated head motion parameters were also included in the GLM to remove the effects of head motion.

In the group-level analyses of GLM 1, paired t-tests were conducted on the contrasts of *Self_HonCon_Lie* versus *Charity_HonCon_Lie* and *Self_NoHonCon_Truth* versus *Charity_NoHonCon_Truth*. The contrast of *Self_HonCon_Lie* versus *Charity_HonCon_Lie* corresponds to the altruistic impact on the decision to lie. The contrast of *Self_NoHonCon_Truth* versus *Charity_NoHonCon_Truth* corresponds to the altruistic impact on the decision to tell the truth in the control condition. To investigate the neural representation of individual differences in the altruistic impact on lying, the estimated altruistic impact on honesty concerns of each participant (i.e., $HC_{PL:\ Self\text{-}Charity}$) was entered into a group-level regression analysis examining the neural response of altruistic impact on lying (*Self_HonCon_Lie* versus *Charity_HonCon_Lie*). In GLM 1, due to insufficient trials (<5) in at least one of the regressors of interest, data from 5 participants were removed from the analysis and data from the remaining 37 participants were analyzed.

In GLM 1, payoff differences between the condition without honesty concerns and the condition with honesty concerns might cause confounds. To further control the payoff

differences between the two conditions, GLM 2 was applied. Similar to GLM 1, four regressors of interest were defined. More importantly, trials with unbalanced payoffs in the *HonCon_Lie* and *NoHonCon_Truth* were excluded. In other words, the included *HonCon_Lie* and *NoHonCon_Truth* trials have identical payoff structures and the same final payoffs within self-profit and charity-profit conditions for each participant. The onsets of the other events (i.e., trials with no response, trials with decisions that lead to low payoff outcomes to senders, and trials with unbalanced payoff structures) were combined into one *other* regressor. Parametric modulators of payoff differences were added to those four regressors of interest. Six estimated head motion parameters were also included in the GLM 2 to remove the effects of head motion.

In the group-level analyses of GLM 2, to check if the findings from GLM 1 can be replicated, the same paired t-tests were conducted on the contrasts of *Self_HonCon_Lie* versus *Charity_HonCon_Lie* and *Self_NoHonCon_Truth* versus *Charity_NoHonCon_Truth* at the group level. To further investigate whether there is a significant correlation between behavioral and neural altruistic impact on honesty concerns after payoff differences were controlled, the estimated altruistic impact on honesty concerns for each participant (i.e., $HC_{PL:\,Self\text{-}Charity}$) was entered into a group-level regression analysis examining the neural response of altruistic impact on lying (i.e., (*Self_HonCon_Lie –Self_NoHonCon_Truth*) versus (*Charity_HonCon_Lie – Charity_NoHonCon_Truth*)). In GLM 2, data from one participant were excluded due to insufficient trials (< 5) after balancing payoff differences between *HonCon_Lie* and *NoHonCon_Truth* in both the charity-profit condition and the self-profit condition. Therefore, data from 36 participants were included.

In GLM 1 and GLM 2, truth-telling in the condition without honesty concerns was used as a control condition/decision. However, truth-telling in the condition with honesty concerns might contain the process of refraining from dishonest gain. The altruistic impact on the truth-telling in the condition with honesty concerns might be different. Therefore, GLM 3 was applied to investigate the neural basis of the altruistic

impact on both lying and truth-telling in the condition with honesty concerns. Four regressors of interest were defined in GLM 3, namely the onsets of: (1) *Charity_HonCon_Lie*, (2) *Charity_HonCon_Truth*, (3) *Self_HonCon_Lie*, and (4) *Self_HonCon_Truth*. The onsets of the other events (i.e., trials with no response and trials without honesty concerns) were regarded as variables of no interest and combined into one other regressor. Those four regressors of interest were parametrically modulated by the payoff differences between the high payoff and the low payoff for the sender. In the group-level analyses, paired t-tests were conducted on the contrasts of *Self_HonCon_Lie* versus *Charity_HonCon_Lie* and *Self_HonCon_Truth* versus *Charity_HonCon_Truth* at the group level. In GLM 3, only 23 participants had sufficient trials (> 5) in each condition of interest and data from 23 participants were included.

Given the important role of the anterior insula in individual differences in lying behavior (Baumgartner *et al.*, 2013) and in avoiding a loss to a charity (Greening *et al.*, 2014), regression analyses were performed with a priori region of interest (ROI) on the bilateral AI. The ROIs were defined by applying the restriction of $y > 0$ (MNI coordinate) to the left and the right insula anatomical masks, which were defined via Wake Forest University Pickatlas toolbox (Maldjian *et al.*, 2003). Results were considered significant if they survived the threshold of cluster- or voxel-level $P < 0.05$, family-wise error (FWE) corrected within the defined region of interest in the AI (i.e., small volume correction). For other analyses, if there is no additional statement, regions were considered significant if they passed the whole-brain cluster FWE correction at $P < 0.05$, with an uncorrected voxel-level cluster-defining threshold of $P < 0.001$ (Eklund *et al.*, 2016).

6.3. Results

6.3.1. Behavioral results

Participants lied more often to get higher payoffs for charitable organizations (*Charity_HonCon*; mean ± s.d.: 61.94% ± 32.27) than for themselves (*Self_HonCon*; mean ± s.d: 55.08% ± 31.42) (two-sided Wilcoxon signed-rank test, P = 0.02; N = 42). The difference in the percentages of choosing higher payoffs between the *Charity_NoHonCon* condition (mean ± s.d: 93.03% ± 15.84) and the *Self_NoHonCon* condition (mean ± s.d: 89.38% ± 18.16) was marginally significant (two-sided Wilcoxon signed-rank test, P = 0.08).

The results hold true in the sample of 37 participants whose data were used in GLM 1 (Figure 6.2). They lied more often to get higher payoffs for charitable organizations than for themselves (*Charity_HonCon*: 69.06% ± 27.10; *Self_HonCon*: 61.56% ± 27.28; two-sided Wilcoxon signed-rank test, P = 0.02). The difference in the percentages of choosing higher payoffs between two conditions without honesty concerns was marginally significant (two-sided Wilcoxon signed-rank test, P = 0.06; *Charity_NoHonCon*: 92.14% ± 16.70; *Self_NoHonCon*: 88.00% ± 18.95).

The means and the standard deviations of reaction times in four conditions in 37 participants whose data were used in fMRI data analyses (GLM 1) are listed in *Table 6.2*. The difference in reaction times between the *Self_HonCon_Lie* and the *Charity_HonCon_Lie* conditions was marginally significant (two-sided Wilcoxon signed-rank test, P = 0.09). Participants were significantly faster when telling the truth in the *Self_NoHonCon* condition than when they were telling the truth in the *Charity_NoHonCon* condition (two-sided Wilcoxon signed-rank test, P = 0.006).

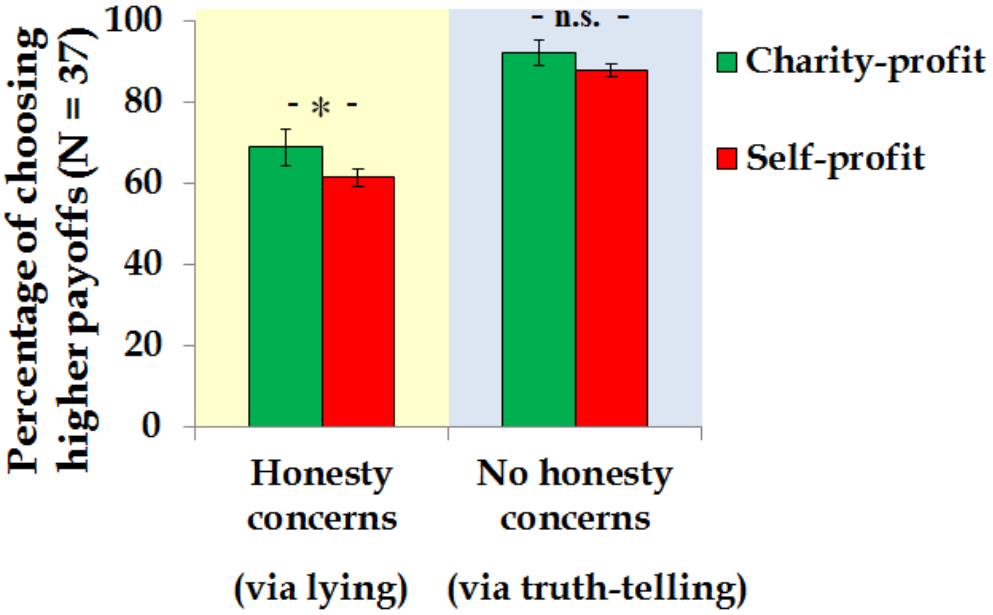

Figure 6.2: The behavioral results in Study 2. The percentages of choosing higher payoffs in four conditions (*Charity_HonCon*, *Self_HonCon*, *Charity_NoHonCon*, and *Self_NoHonCon*) are shown. The condition with honesty concerns refers to the concerns of earning higher payoffs through lying (marked in light yellow; *HonCon*). The condition without honesty concerns refers to the absence of the concerns of earning higher payoffs through lying (marked in light blue; *NoHonCon*). (* $P < 0.05$; n.s.: not significant; N = 37; error bars: s.e.m.)

Table 6.2
Reaction time data in Study 2 (ms; N = 37)

Beneficiary	**Mean (s.d.)**	
	HonCon_Lie	***NoHonCon_Truth***
Charity	1577 (531)	1367 (414)
Self	1538 (531)	1312 (419)

HonCon_Lie: lying in the conditions with honesty concerns
NoHonCon_Truth: truth-telling in the conditions without honesty concerns

6.3.2. Functional MRI results

In GLM 1, the contrast corresponding to the altruistic impact on lying (i.e., *Self_HonCon_Lie* versus *Charity_HonCon_Lie*) showed significant activation in the right AI (Figure 6.3; Table 6.3). The left AI was also activated under a more lenient threshold of P < 0.001 uncorrected (k = 88; MNI coordinate of the peak voxel: -28, 20, -8). No significant results were observed in the opposite contrast. No significant results were observed in the contrast corresponding to the altruistic impact on truth-telling in the control condition (i.e., *Self_NoHonCon_Truth* versus *Charity_NoHonCon_Truth*), even with a lenient threshold (i.e., uncorrected voxel-level threshold of P < 0.005). The results of regression analyses showed significant positive correlations (small volume correction, $P_{FWE\text{-}corrected} < 0.05$) between the altruistic impact on activity in the left and right AI (MNI coordinate of the peak voxel: -28, 12, -12; 32, 24, 6; *Self_HonCon_Lie* versus *Charity_HonCon_Lie*) and the estimated altruistic impact on honesty concerns (i.e., $HC_{PL:\ Self\text{-}Charity}$).

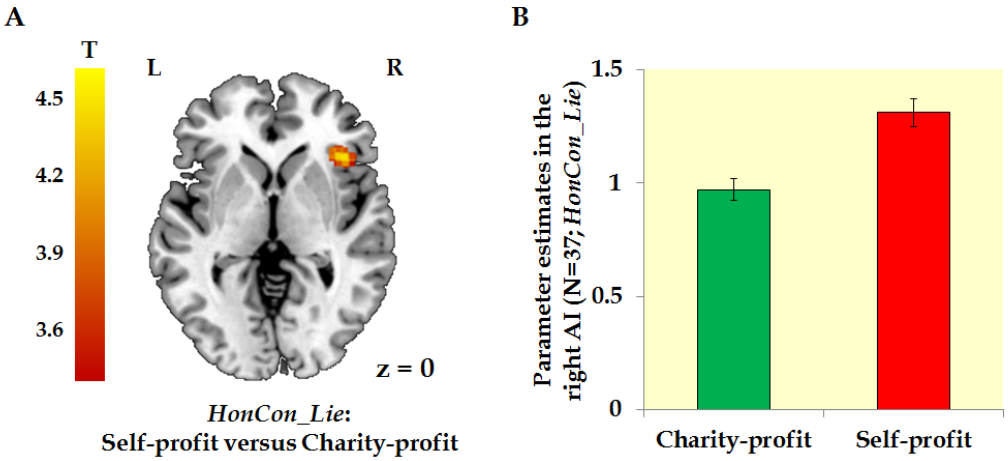

Figure 6.3: fMRI results of GLM 1 in Study 2. The results of the contrast of *Self_HonCon_Lie* versus *Charity_HonCon_Lie* are shown (N = 37). (A) Significant activation was observed in the right AI. (B) Parameter estimates were extracted from the whole activated cluster in the right AI in the *Charity_HonCon_Lie* and the *Self_HonCon_Lie* conditions. (*HonCon_Lie*: lying in the conditions with honesty concerns; AI: anterior insula; error bars: s.e.m.)

In GLM 2, after controlling for payoff differences, the altruistic impact on lying was related to activity in the right AI (*Self_HonCon_Lie* versus *Charity_HonCon_Lie*) under a lenient threshold (MNI coordinate of the peak voxel: 36, 24, -4; voxel-level P < 0.001 uncorrected and k = 133). No significant activation in the AI was observed in the opposite contrast. No significant results were observed in the comparison between *Self_NoHonCon_Truth* versus *Charity_NoHonCon_Truth*. Significant positive correlations (small volume correction, $P_{FWE\text{-}corrected} < 0.05$) were also observed between the altruistic impact on lying-specific activity in the left and right AI (MNI coordinate of the peak voxel: -36, 18, 2; 34, 20, 4; (*Self_HonCon_Lie* – *Self_NoHonCon_Truth*) versus (*Charity_HonCon_Lie* – *Charity_NoHonCon_Truth*)) and the estimated altruistic impact on honesty concerns ($HC_{PL:\ Self\text{-}Charity}$; Figure 6.4).

Table 6.3
fMRI results of GLM 1 in Study 2 (N = 37)

Brain area	L/R	Voxel	MNI coordinates			T value
			x	y	z	
Charity_HonCon_Lie > Self_HonCon_Lie						
None						
Self_HonCon_Lie > Charity_HonCon_Lie						
Anterior insula	R	247	36	24	-4	4.66
Charity_NoHonCon_Truth > Self_NoHonCon_Truth						
None						
Self_NoHonCon_Truth > Charity_NoHonCon_Truth						
None						

Voxel-level threshold P < 0.001 uncorrected, cluster-level $P_{FWE\text{-}corrected} < 0.05$.
Charity_HonCon_Lie: lying in the condition with honesty concerns and with a charity as the beneficiary.
Self_HonCon_Lie: lying in the condition with honesty concerns and with participants as the beneficiary.
Charity_NoHonCon_Truth: truth-telling in the condition without honesty concerns and with a charity as the beneficiary.
Self_NoHonCon_Truth: truth-telling in the condition without honesty concerns and with participants as the beneficiary.

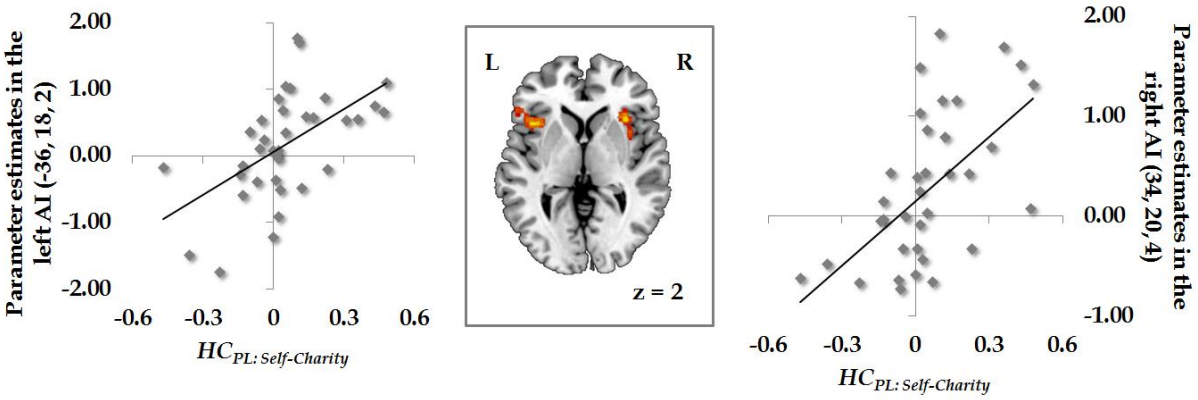

Figure 6.4: fMRI results of GLM 2 in Study 2. The neuroimaging results of the impact of altruistic outcomes on honesty concerns are shown (N = 36). Neural activity in the left and the right AI in the contrast of (*Self_HonCon_Lie –Self_NoHonCon_Truth*) versus (*Charity_HonCon_Lie – Charity_NoHonCon_Truth*) positively correlated with $HC_{PL:\ Self\text{-}Charity}$. All effects were significant after small volume correction ($P_{FWE\text{-}corrected} < 0.05$). For illustration purpose, activations in the AI are displayed at uncorrected significance threshold ($P < 0.005$, $k > 100$). (AI: anterior insula; $HC_{PL:\ Self\text{-}Charity}$: the difference in the ratio of the payoff loss caused by honesty concerns between the self-profit condition and the charity-profit condition)

In GLM 3, similar to previous findings in GLM 1 and GLM 2, significant activation in the right AI was observed in the contrast of *Self_HonCon_Lie* versus *Charity_HonCon_Lie* (Figure 6.5A; Table 6.4). Truth-telling to refrain from dishonest gain for charitable organizations elicited higher activity in the ventral medial prefrontal cortex (VMPFC) than truth-telling to refrain from dishonest gain for participants themselves (*Charity_HonCon_Truth* versus *Self_HonCon_Truth*; Figure 6.5B; Table 6.4). No significant results were observed in the opposite contrasts (Table 6.4).

Table 6.4
fMRI results of GLM 3 in Study 2 (N = 23)

Brain area	L/M/R	Voxel	MNI coordinates			T value
			x	y	z	
Charity_HonCon_Lie > Self_HonCon_Lie						
None						
Self_HonCon_Lie > Charity_HonCon_Lie						
Anterior insula	R	273	34	24	-6	5.31
Charity_HonCon_Truth > Self_HonCon_Truth						
Ventral medial prefrontal cortex	M	202	-2	42	-18	4.73
Self_HonCon_Truth > Charity_HonCon_Truth						
None						

Voxel-level threshold $P < 0.001$ uncorrected, cluster-level $P_{FWE\text{-}corrected} < 0.05$.
Charity_HonCon_Lie: lying in the condition with honesty concerns and with a charity as the beneficiary.
Self_HonCon_Lie: lying in the condition with honesty concerns and with participants as the beneficiary.
Charity_HonCon_Truth: truth-telling in the condition with honesty concerns and with a charity as the beneficiary.
Self_HonCon_Truth: truth-telling in the condition with honesty concerns and with participants as the beneficiary.

6.4. Discussion

A modified sender-receiver game was used in Study 2 to investigate how altruistic outcomes modulate lying behaviors and the underlying neural processes from the perspective of the communicators. The altruistic outcome was introduced by including charitable benefits in the decision-making process. In the charity-profit condition, lies could benefit a charity, whereas in the self-profit condition lies could only benefit the communicator. In the control conditions, participants could achieve greater benefits for a charity or themselves by telling the truth. Consistent with previous behavioral findings (Lewis *et al.*, 2012), the behavioral results in Study 2 showed that participants lied more often to earn a higher payoff for a charity than for themselves.

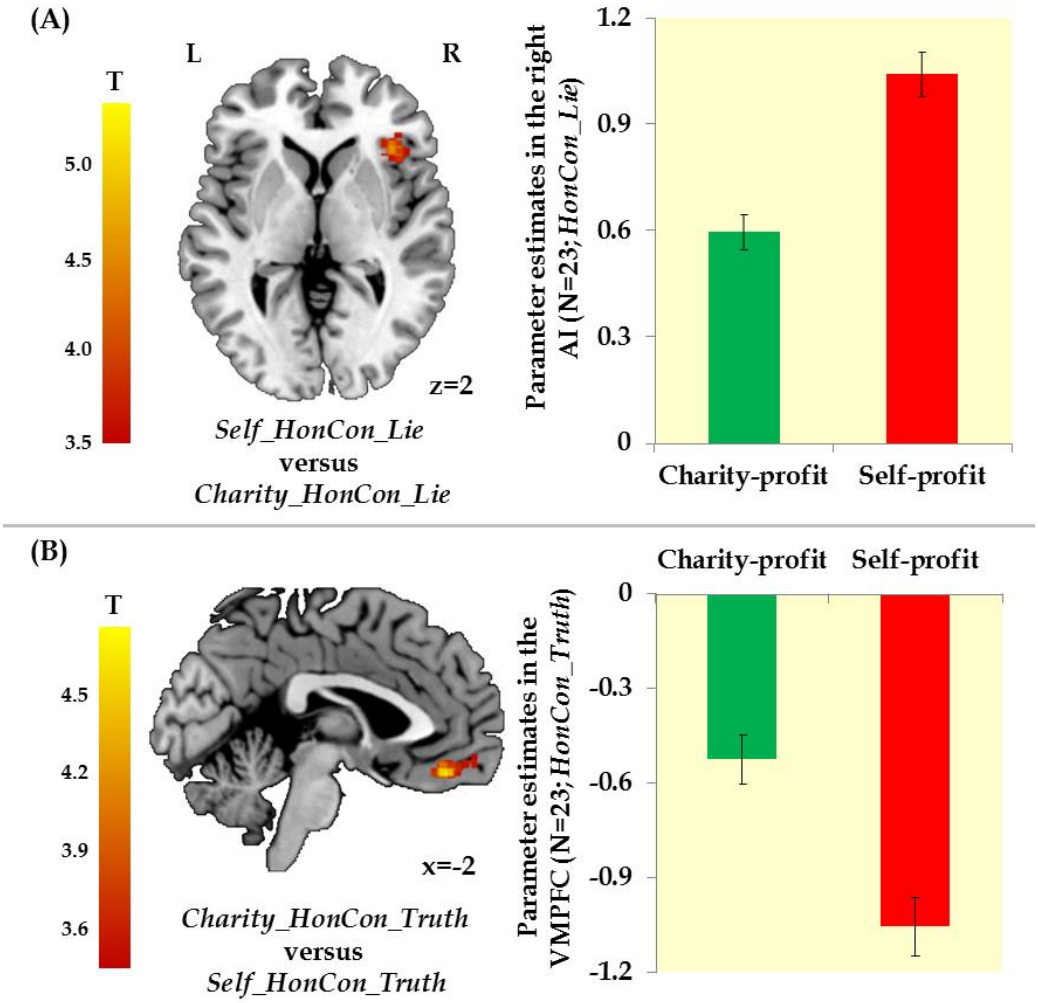

Figure 6.5: fMRI results of GLM 3 in Study 2 (N = 23). (A) The results of the contrast of *Self_HonCon_Lie* versus *Charity_HonCon_Lie* are shown. Significant activation was observed in the right AI. Parameter estimates were extracted from the whole activated cluster in the right AI in the *Charity_HonCon_Lie* and the *Self_HonCon_Lie* conditions. (B) The results of the contrast of *Charity_HonCon_Truth* versus *Self_HonCon_Truth* are shown. Significant activation was observed in the VMPFC. Parameter estimates were extracted from the whole activated cluster in the VMPFC in the *Charity_HonCon_Truth* and the *Self_HonCon_Truth* conditions. (*Self_HonCon_Lie*: lying in the conditions with honesty concerns and with participants as the beneficiary; *Charity_HonCon_Lie*: lying in the conditions with honesty concerns and with a charity as the beneficiary; *Self_HonCon_Truth*: truth-telling in the conditions with honesty concerns and with participants as the beneficiary; *Charity_HonCon_Truth*: truth-telling in the conditions with honesty concerns and with a charity as the beneficiary; AI: anterior insula; VMPFC: ventral medial prefrontal cortex; error bars: s.e.m.)

At the neural level, the results in GLM 1 showed that lying for a charity reduced AI activity compared with lying for oneself. The AI is one of the key regions associated with deception (Baumgartner *et al.*, 2009; Christ *et al.*, 2009; Baumgartner *et al.*, 2013; Farah *et al.*, 2014; Lisofsky *et al.*, 2014; Sun *et al.*, 2015a; Volz *et al.*, 2015). In addition to deception, the insula signals negative emotional states (Calder *et al.*, 2000) or aversive emotional experiences, such as the experience of unfairness (Sanfey *et al.*, 2003; Tabibnia *et al.*, 2008) and the threat of punishment (Spitzer *et al.*, 2007). The consistent findings on the involvement of the AI in deceptive decision-making might be due to the negative emotional response to lying.

In an fMRI study of charitable donations, decisions to oppose donations to charities were associated with activity in the lateral orbitofrontal cortex and the AI (Moll *et al.*, 2006), suggesting a role for the AI in mediating aversive experiences (Moll *et al.*, 2008). In a study investigating the neural response to aversive drinks that were intentionally delivered by others, participants felt more angry toward the intentional aversive conditions and their AI activity correlated with the interaction between the perceived intentionality and anticipated outcome valence (Liljeholm *et al.*, 2014). Given that the AI is sensitive to aversive social interactions (Rilling and Sanfey, 2011) and the interference of intentionality, reduced AI activity observed when participants lied for a charity in Study 2 might be caused by the process in which altruistic outcomes modulated the negative affective states when participants chose to lie.

In a recent resting EEG study, people with the higher neural baseline activation in the AI had a lower propensity to deceive (Baumgartner *et al.*, 2013), suggesting a potential link between the AI and individual differences in the reluctance to lying. In Study 2, participants whose concerns regarding honesty were more strongly decreased by the altruistic outcomes (i.e., with larger $HC_{PL:\ Self\text{-}Charity}$) showed greater reductions in lying-specific activity in the bilateral AI in the charity-profit condition. The result further supports the hypothesis that the anterior insula signals the aversive emotional responses to lying, which can further be modulated by the outcomes (self-profit or altruistic outcomes). In a study of charitable decision-making, increased

connectivity between the VMPFC and the bilateral AI was observed in the free and forced donation, suggesting that the AI encodes social values of donation (Hare *et al.*, 2010). A recent study of social exclusion showed that the bilateral AI was activated when participants observed others being excluded, whereas participants with stronger neural activity in the right AI in response to other's social pain behaved in a more prosocial manner toward the victims (Masten *et al.*, 2011). Given the above findings and the role of AI in empathy (Singer *et al.*, 2004b; Jackson *et al.*, 2005; Lamm *et al.*, 2011), a precursor of altruistic behavior (Singer and Lamm, 2009), the right AI might code for the modulation of altruistic outcomes when making aversive decisions.

Intriguingly, the involvement of the DLPFC was not observed in Study 2. The DLPFC activation was observed in tasks that require high level of cognitive control (Sanfey *et al.*, 2003; Aron *et al.*, 2004; Spitzer *et al.*, 2007), particularly in deceptive decision-making tasks (Lee *et al.*, 2002; Lee *et al.*, 2005; Luan Phan *et al.*, 2005; Nunez *et al.*, 2005; Abe *et al.*, 2006; Abe *et al.*, 2007; Karton and Bachmann, 2011; Karton *et al.*, 2014). Zhu et al. (2014) found that the DLPFC lesioned patients showed reduced honesty concerns in the sender-receiver game. However, the concerns for receivers' payoffs and the confounding truth-telling behaviors (i.e., sophisticated deception) were controlled for in Study 2. It is possible that the cognitive resources required in the self-profit and charity-profit conditions when people are making deceptive decisions are similar. The altruistic goals might modulate honesty concerns by influencing the affective processing via the AI, rather than by affecting the cognitive control via the DLPFC.

When participants decided to tell the truth and reduce payoffs for a charity, activity in the VMPFC was significantly higher than when they decided to tell the truth and reduce payoffs for themselves. In a previous study, participants were more likely to donate to a charity with a high subjective value and the VMPFC was involved in computing the value of charitable donations (Hare et al., 2010). Moreover, VMPFC activation was observed in the comparison between costly decisions (either costly donations or costly opposition to donations) and pure monetary rewards (Moll et al.,

2006). In Study 2, participants chose a preferable charitable organization to be the beneficiary of their decisions before the fMRI experiment. VMPFC activation observed here might not reflect the values of different charitable organizations, but the VMPFC might be involved in the process of sacrificing social rewards to avoid the psychological costs of lying.

6.5. Limitations

The association between the AI and lying aversion might suffer the problem of reverse inference (i.e., inferring a cognitive process from neuroimaging data) (Poldrack, 2006). Trust in the reverse inference can be enhanced by additional measures such as measures of mood (Fehr, 2009). However, in Study 2, participants were not asked to report emotional valences. Nevertheless, the measurement of lying frequency and the estimated altruistic impact on honesty concerns support the speculation that participants were generally less reluctant and averse to lie for a charity.

7. Study 3: The neural mechanisms underlying the modulation of altruistic outcomes on recipients' process of deception

The associated work has been published in Yin Lijun and Weber Bernd (2016). Can beneficial ends justify lying? Neural responses to the passive reception of lies and truth-telling with beneficial and harmful monetary outcomes. Social Cognitive and Affective Neuroscience (11), 423-432.

Author contributions: L.Y. and B.W. designed research; L.Y. performed research; L.Y. analyzed data; L.Y. and B.W. wrote the paper.

7.1. Introduction

The modulation of altruistic outcomes on communicators' process of deception was investigated in Study 2. The next question is about the altruistic impact on the recipients' neural process of deception. By showing participants the scenarios that contain different social contexts and by asking them to provide moral judgments in the scanner, researchers could investigate the neural correlates of evaluations of lying and truth-telling (Berthoz *et al.*, 2006; Hayashi *et al.*, 2010; Wu *et al.*, 2011; Hayashi *et al.*, 2014). However, three limitations exist if the method is used to investigate the proposed research question. First, participants' affective or emotional responses to the scenarios might be interrupted if they are asked to provide moral judgments during the experiment (Knutson *et al.*, 2014). Second, making judgments about a given scenario from a third-party perspective might involve rational processing (Wu *et al.*, 2011) and is less emotionally arousing. The rational process might further induce greater involvement of the prefrontal cortex (Abe *et al.*, 2014; Hayashi *et al.*, 2014). Finally, different neural correlates are involved in situations where oneself or

others are deceived. For example, the amygdala was activated when participants judged themselves as being deceived compared with the judgments about others being deceived, participants deceiving the experimenter, and others deceiving the experimenter (Grezes *et al.*, 2006).

Considering the confounds and personal involvement, a modified sender-receiver game (Erat and Gneezy, 2012) was used in Study 3 to investigate the neural processes underlying lies from the perspectives of the recipients. In this game, participants were the direct recipients of lies and truth, with different monetary outcomes. Participants acting as receivers first read the messages, which were sent by multiple senders, regarding the results of a die roll representing one of the payoff options. They then decided whether to believe the message or not. If participants did not believe the messages, both players received the minimum payoffs. If participants believed senders' messages, they would face four different conditions: the *beneficial lies condition:* senders sent untruthful messages and both players earned more money compared to the alternative options; the *harmful lies condition:* senders sent untruthful messages and the senders earned more money but participants earned less money; the *beneficial truth condition:* senders sent truthful messages and both players earned more money; and the *harmful truth condition:* senders sent truthful messages and the senders earned more money but participants earned less money. After the fMRI experiment, an explicit moral judgment task was used to measure participants' judgments about the moral acceptance of the different conditions.

7.2. Materials and methods

7.2.1. Participants

Forty-one participants (23 females; mean ± s.d. age = 24.15 ± 3.14 years ranged from 19 to 32 years) were invited to take part in the fMRI experiment and play the game as receivers. All participants reported no prior history of psychiatric or neurological

disorders and had normal or corrected-to-normal vision. All participants provided their informed consent. The study was approved by the Ethics Committee of the University of Bonn.

7.2.2. Tasks

A modified sender-receiver game (Figure 7.1) was used. In every trial, there were two payoff options. Each consisted of the payoff for a sender (represented by the blue bar) and the payoff for a receiver (represented by the red bar). The outcome of a die roll represented one of two payoff options (indexed by the icon of a die). The senders chose one of six numbers to phrase a message: "The outcome of the die roll is x" (where x = 1, 2, 3, 4, 5, or 6). The senders could either send an untruthful message (Figure 7.1A) or a truthful message (Figure 7.1B) to the receivers.

In the scanner, participants as the message receivers first saw the message which indicated the outcome of the die roll and the initials of the associated sender. On the same screen, participants made the decision to believe (i.e., choosing "Yes") or not to believe (i.e., choosing "No") the sender's message by pressing the respective button on the response grips. The positions of "Yes" and "No" options were counterbalanced within each participant. After pressing the button on the response grips, a yellow frame appeared and indexed the corresponding choice for 0.5s. A fixation cross was then displayed for a jittered interval (4-6s). If participants believed the sender's message, the next screen displayed the two payoff options and the actual outcome of the die roll for 5s. Participants could find out that the sender lied or told the truth. If the message was untruthful, the alternative option would be implemented (Figure 7.1C). Otherwise, the payoff option represented by the die would be implemented (Figure 7.1D). The yellow frame indexed the final implemented payoff option. If participants did not believe, both sides of the players earned a minimum payoff (i.e., 1€; Figure 7.1E).

In the original sender-receiver game from the study by Erat and Gneezy (2012), if participants chose the actual outcome of a die roll, the option represented by the die

would be implemented. Otherwise, the alternative option would be implemented. When participants were making decisions in the original sender-receiver game, extra cognitive processes of generating strategies or emotional responses toward bad choices might be involved. These might influence participants' neural evaluations of senders' behaviors. To avoid these confounding factors, in the task of Study 3, if a receiver did not believe the sender's message, both sides of the players earned a minimum payoff.

7.2.3. Design and stimuli

The current event-related fMRI study adopted a 2 (mean: lies or truth) by 2 (outcome: beneficial or harmful outcomes) within-subject design. One 40-min scanning run contained 144 trials in total (36 trials per condition). Before the fMRI experiment, 86 participants were invited to play the game online as senders. They made decisions in 144 trials and earned 5€ for their participation. Their messages were pre-selected to phrase messages for all four conditions in the fMRI experiment, and the initials of the senders would not repeat more than three times. The payoff structure is shown in Table 7.1. The final payoff for both the sender and the receiver was one of three monetary amounts (4€, 8€, or 12€). The alternative payoff for the sender was decreased by 25% or 75% compared to the payoff in the implemented option. The alternative payoff for the receiver was increased or decreased by 25% or 75% compared to the payoff in the implemented option.

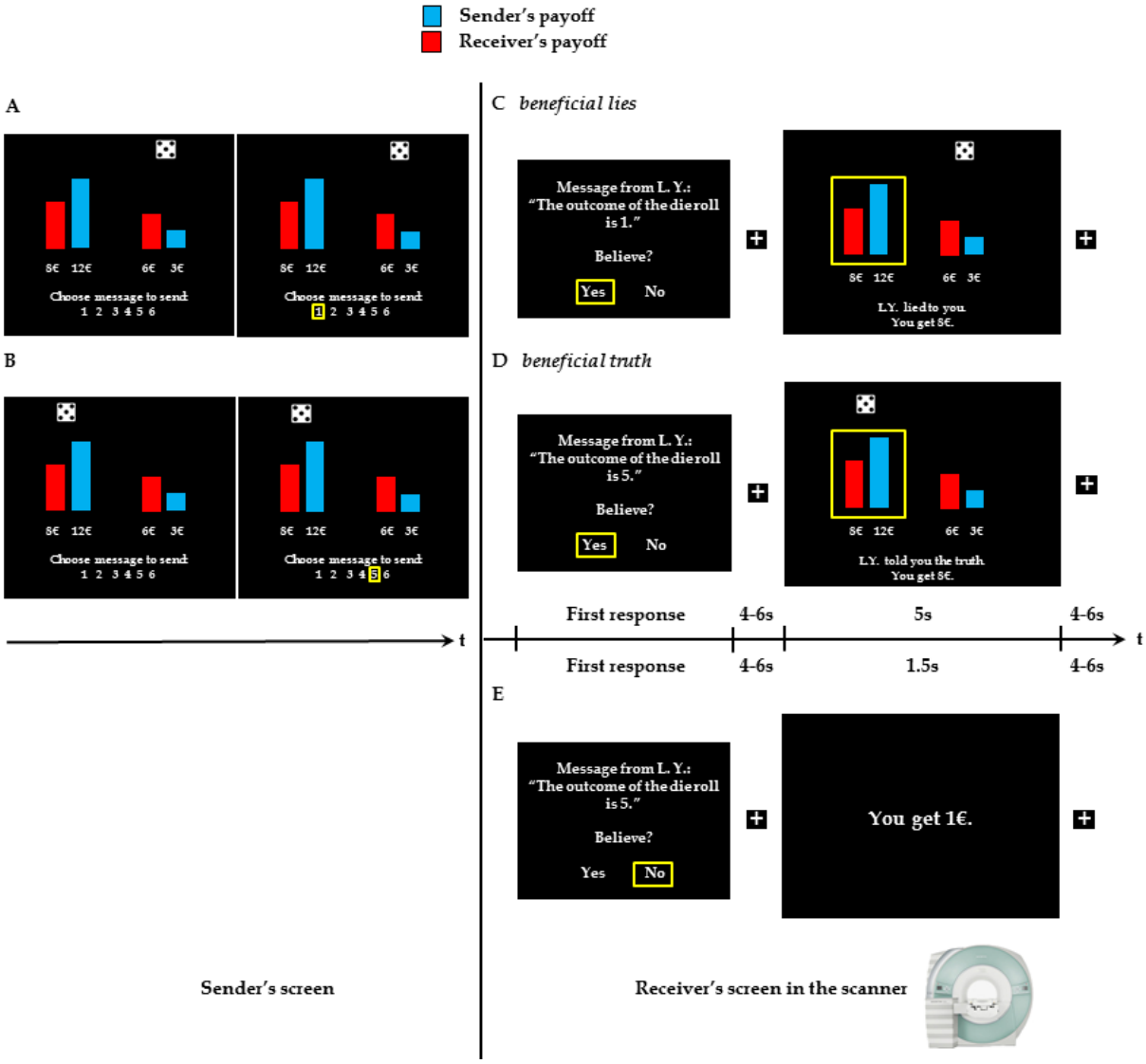

Figure 7.1: The experimental paradigm of the modified sender-receiver game in Study 3. (A) The outcome of the die roll represented one of the payoff options (e.g., "5" represented the payoff option on the right, i.e., 6€ for a receiver and 3€ for a sender). A sender (e.g., L.Y.) sent a message to a receiver (e.g., the untruthful message of "The outcome of the die roll is 1"). (B) The outcome of the die roll represented the payoff option on the left (8€ for a receiver and 12€ for a sender). The sender sent a truthful message of "The outcome of the die roll is 5." (C) In the scanner, if a participant believed the sender's untruthful message, the alternative payoff option would be implemented (the option within the yellow outlined frame). (D) If the participant believed the sender's truthful message, the payment option represented by the die would be implemented. (E) If the participant did not believe, both the participant and the sender earned 1€.

7.2.4. Procedure

Before scanning, fMRI participants (i.e., receivers) completed a questionnaire to ensure that they fully understood the experiment. The senders and the receivers were told that their identity remained confidential to the opponents, and the initials of the senders would be displayed to the receivers. Participants were informed that the senders knew all information before they sent the messages, including the two payoff options and the results of the die rolls. In the scanner, participants received the messages and made their decisions to believe or not.

After the fMRI experiment, participants were asked to rate emotional valences of four experimental conditions according to Lang's Self-Assessment-Manikin Valence Scale (Lang, 1980). The nine-level Self-Assessment Manikin valence scale (1 = very unhappy, 5 = neutral, 9 = very happy) was adapted from PXLab (Irtel, 2008). They were also asked to rate moral acceptance of four conditions according to a nine-level scale (1 = not morally acceptable at all, 5 = neutral, 9 = extremely morally acceptable). One of the trials was randomly selected. The senders and the receivers were paid accordingly. In addition to the payoffs from the game, the receivers earned an extra 10€ for their participation in the fMRI experiment.

7.2.5. Data acquisition

Participants' responses in the scanner were collected via an MRI-compatible response device (NordicNeuroLab, Bergen, Norway). All images were acquired on a Siemens Trio 3.0-Tesla scanner with a standard 32-channel head coil. Structural scans included T1-weighted images (TR = 1660 ms; TE = 2.75 ms; flip angle = 9°; slice thickness = 0.8 mm). One functional session was run, starting with a localizer scan, and was then followed by the paradigm implemented in Presentation (Neurobehavioral Systems; http://www.neurobs.com) during which T2*-weighted echo planar images were collected (flip angle = 90°; TR = 2500 ms; TE = 30 ms; 96 ×

96 acquisition matrix; field of view =192 mm × 192 mm; 37 slices; in-plane resolution = 2 × 2 mm², thickness = 3 mm).

Table 7.1
The payoff structure in Study 3

Ratios S/R / Alternatives S/R / Final payoffs S/R	Beneficial condition				Harmful condition			
	-25%/ -25%	-75%/ -75%	-75%/ -25%	-25%/ -75%	-25%/ +25%	-75%/ +75%	-75%/ +25%	-25%/ +75%
4/4	3/3	1/1	1/3	3/1	3/5	1/7	1/5	3/7
8/4	6/3	2/1	2/3	6/1	6/5	2/7	2/5	6/7
12/4	9/3	3/1	3/3	9/1	9/5	3/7	3/5	9/7
4/8	3/6	1/2	1/6	3/2	3/10	1/14	1/10	3/14
8/8	6/6	2/2	2/6	6/2	6/10	2/14	2/10	6/14
12/8	9/6	3/2	3/6	9/2	9/10	3/14	3/10	9/14
4/12	3/9	1/3	1/9	3/3	3/15	1/21	1/15	3/21
8/12	6/9	2/3	2/9	6/3	6/15	2/21	2/15	6/21
12/12	9/9	3/3	3/9	9/3	9/15	3/21	3/15	9/21

S: Sender; R: Receiver.
Negative ratio: the alternative payoff was decreased compared to the final payoff (beneficial conditions).
Positive ratio: the alternative payoff was increased compared to the final payoff (harmful conditions).

7.2.6. Data analyses

Data from three subjects were excluded due to excessive head movements (i.e., > 3 mm or 3° of rotation). The data from the remaining 38 participants (22 females; mean ± s.d = 24 ± 3.24 years ranged from 19 to 32 years) were analyzed.

7.2.6.1. Behavioral analyses

Statistical analyses of ratings of emotional valences and moral acceptance were conducted with SPSS 22.0 (IBM Corporation, Armonk, NY, USA). Repeated-measure

analysis of variance (ANOVA) models were performed. All reported P values were two-tailed, and $P < 0.05$ was considered statistically significant. Post hoc analysis with Bonferroni correction was applied for significant interaction effects if any.

7.2.6.2. Functional MRI data analyses

SPM8 was adopted for fMRI data analyses (Welcome Department of Cognitive Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm/). For each participant, EPI images were first realigned and resliced. The anatomical image was then co-registered with the mean EPI image of each participant which was further segmented. The SPM8's DARTEL tool was used to create a template and normalize functional and anatomical scans to the MNI template. Finally, the normalized functional images were smoothed by an 8-mm FWHM Gaussian filter. High-pass temporal filtering with a cut-off of 128s was performed to remove low-frequency drifts.

Based on the events of interest, trials were categorized as the following conditions: *beneficial lies* (the untruthful messages that made the receiver earn more money; Figure 7.1C), *beneficial truth* (the truthful messages that made the receiver earn more money; Figure 7.1D), *harmful lies* (the untruthful messages that made the receiver earn less money), and *harmful truth* (the truthful messages that made the receiver earn less money). Statistical analyses of fMRI data were estimated using a general linear model (GLM). The regressors of interest included the onsets of the outcome phases in the following conditions: (1) *beneficial lies*, (2) *beneficial truth*, (3) *harmful lies*, and (4) *harmful truth*. Onsets of the other events (i.e., the decision phases, trials with no response, and trials in which participants did not believe senders' messages; Figure 7.1E) were combined into one *other* regressor of no interest. Six estimated head motion parameters were also included in the GLM to account for the residual effects of head motion.

In the group-level analyses, a flexible factorial model with two within-group factors (outcome (beneficial or harmful outcomes) and mean (lying or truth-telling)) was performed. If there is no additional statement, regions were considered significant if

they passed the whole-brain cluster FWE correction at $P < 0.05$, with an uncorrected voxel-level cluster-defining threshold of $P < 0.001$ (Eklund *et al.*, 2016). Given the important role of the amygdala in judgments of deceit (Grezes *et al.*, 2006), interaction analyses were performed with a priori region of interest (ROI) on the amygdala. The ROIs were defined by using the left and the right amygdala anatomical masks in Wake Forest University Pickatlas (WFU) toolbox (Maldjian *et al.*, 2003). Results were considered significant if they survived the threshold of cluster-level $P < 0.05$, family-wise error (FWE) corrected within the defined region of interest in the amygdala (i.e., small volume correction).

7.3. Results

7.3.1. Behavioral results

The mean ratings and the standard deviations of emotional valences and moral acceptance are listed in Table 7.2.

Table 7.2
The behavioral ratings in Study 3 (N = 38; mean ± s.d.)

Ratings	Outcomes	Means	
		Truth	Lies
Valence	**Beneficial outcome**	7.79 (1.43)	7.05 (1.73)
	Harmful outcome	5.55 (1.63)	4.24 (1.85)
Moral acceptance	**Beneficial outcome**	8.34 (1.12)	6.97 (2.03)
	Harmful outcome	6.61 (2.04)	4.58 (2.16)

Valence: 1 = very unhappy, 5 = neutral, 9 = very happy.
Moral acceptance: 1 = not morally acceptable at all, 5 = neutral, 9 = extremely morally acceptable.

With respect to emotional valences, a 2 (mean: lying or truth-telling) by 2 (outcome: beneficial or harmful outcomes) ANOVA was performed (Figure 7.2A). The main effects of mean and outcome were significant (truth-telling > lying: $F_{(1, 37)} = 80.50$, $P <$

0.001; beneficial outcomes > harmful outcomes: $F_{(1, 37)} = 15.59$, $P < 0.001$). The interaction effect of mean × outcome was significant ($F_{(1, 37)} = 5.66$, $P = 0.02$). The post hoc analysis showed that the valence difference between *harmful truth* and *harmful lies* was significantly higher than the valence difference between *beneficial truth* and *beneficial lies* ($t_{(37)} = 0.02$; $P = 0.02$).

With respect to moral acceptance ratings, a 2 (mean: lying or truth-telling) by 2 (outcome: beneficial or harmful outcomes) ANOVA was performed (Figure 7.2B). The main effects of mean and outcome were significant (truth-telling > lying: $F_{(1, 37)} = 75.39$, $P < 0.001$; beneficial outcomes > harmful outcomes: $F_{(1, 37)} = 32.34$, $P < 0.001$). The interaction effect of the mean and outcome showed only a trend ($F_{(1, 37)} = 1.81$, $P = 0.08$).

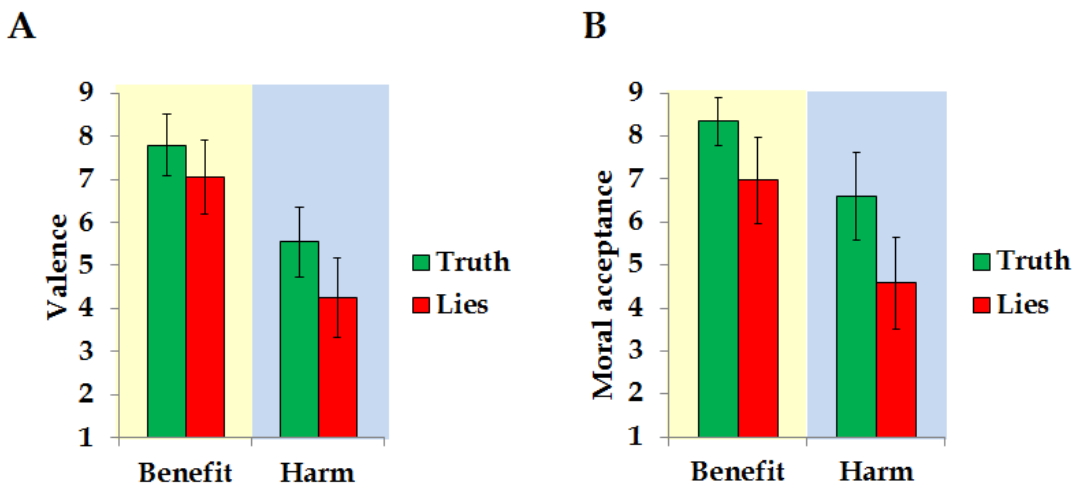

Figure 7.2: The behavioral results in Study 3. Participants' emotional valences (A) and moral acceptance ratings (B) of four conditions (*beneficial truth*, *beneficial lies*, *harmful truth*, and *harmful lies*) are shown (error bars: s.d.).

7.3.2. Functional MRI results

The results are listed in Table 7.3. The main effect of outcome (beneficial outcomes versus harmful outcomes) elicited stronger activation in the bilateral NAcc (Figure 7.3*A*). The opposite comparison yielded no results. The main effect of truth versus lies elicited stronger activation in the left NAcc (Figure 7.3*B*). Figure 7.4*A* reveals the overlapping region (i.e., the left NAcc; marked in yellow) that was activated in the main effects of benefit versus harm (marked in red) and truth versus lies (marked in green). The BOLD percentage signal changes were extracted from the overlapping region in all four conditions (Figure 7.4*B*).

The main effect of mean (lies versus truth; Figure 7.5*A*) activated the left supplementary motor area (SMA), the right inferior frontal gyrus (IFG), the right superior temporal sulcus (STS), and the left anterior insula (AI). The BOLD percentage signal changes were extracted from the SMA (Figure 7.5*B*), the right IFG (Figure 7.5*C*), the right STS (Figure 7.5*D*), and the left AI (Figure 7.5*E*) in all four conditions (i.e., *beneficial truth*, *harmful truth*, *beneficial lies*, and *harmful lies*).

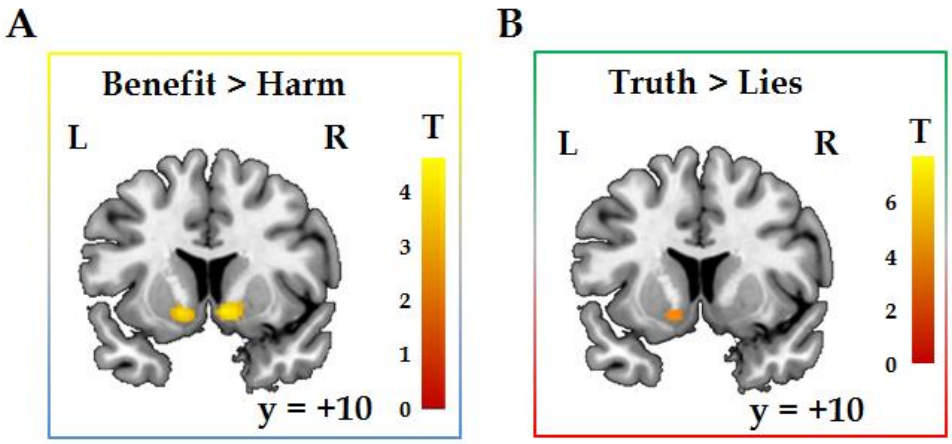

Figure 7.3: fMRI results in Study 3. (A) The significant main effect of beneficial outcomes versus harmful outcomes was observed in the bilateral NAcc. (B) The significant main effect of truth versus lies was observed in the left NAcc. (NAcc: nucleus accumbens)

Table 7.3
fMRI results in Study 3 (N = 38)

Brain area	L/R	Cluster	MNI coordinates			Z value
			x	y	z	
Truth > Lies						
Occipital Lobe	R	17215	12	-65	-3	6.50
Nucleus accumbens	L	374	-12	4	-10	4.20
Lies > Truth						
Superior temporal sulcus	R	878	58	-48	12	4.51
Supplementary motor area	L	372	-3	22	60	4.09
Anterior insula	L	475	-36	22	9	4.08
Inferior frontal gyrus	R	480	40	16	24	4.01
Beneficial outcomes > Harmful outcomes						
Nucleus accumbens	R	725	12	12	-6	4.64
Nucleus accumbens	L	386	-10	7	-7	4.34
Harmful outcomes > Beneficial outcomes						
None						
(Beneficial Truth - Harmful Truth) versus (Beneficial Lies - Harmful Lies)						
Amygdala*	L	52	-17	0	-18	3.52
(Beneficial Lies - Harmful Lies) versus (Beneficial Truth - Harmful Truth)						
None						

Voxel-level threshold $P < 0.001$, cluster-level $P_{FWE\text{-}corrected} < 0.05$.
*The effect is significant after small-volume correction for multiple comparisons within the amygdala ROI ($P_{FWE\text{-}corrected} < 0.05$).

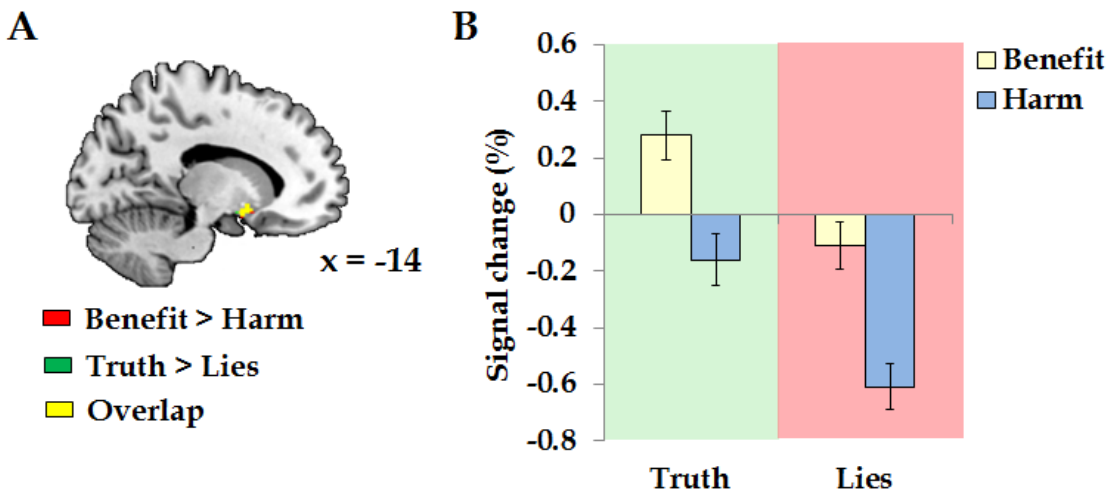

Figure 7.4: fMRI results in Study 3. (A) The overlapping region (i.e., the left NAcc; yellow), which was activated in the contrast of beneficial outcomes versus harmful outcomes (red) and the contrast of truth versus lies (green), is shown (masked with the NAcc anatomical mask from WFU Pickatlas Tool (Maldjian *et al.*, 2003)). (B) Percentage signal changes were extracted from the overlapping region. (NAcc: nucleus accumbens; error bars: s.e.m.)

The interaction effect of (*beneficial truth – harmful truth*) versus (*beneficial lies - harmful lies*) activated the left amygdala (small-volume correction, $P_{FWE\text{-}corrected} < 0.05$; Figure 7.6*A*). Further analyses showed that no significant differences were observed in the contrasts of *harmful lies* versus *harmful truth* and *beneficial lies* versus *harmful lies* (Ps ≥ 0.08). Significant differences were observed in the contrasts of *beneficial truth* versus *beneficial lies* (P < 0.001) and *beneficial truth* versus *harmful truth* (P = 0.02; Figure 7.6*B*). No significant results were observed in the reversed interaction effect of (*beneficial lies - harmful lies*) versus (*beneficial truth – harmful truth*).

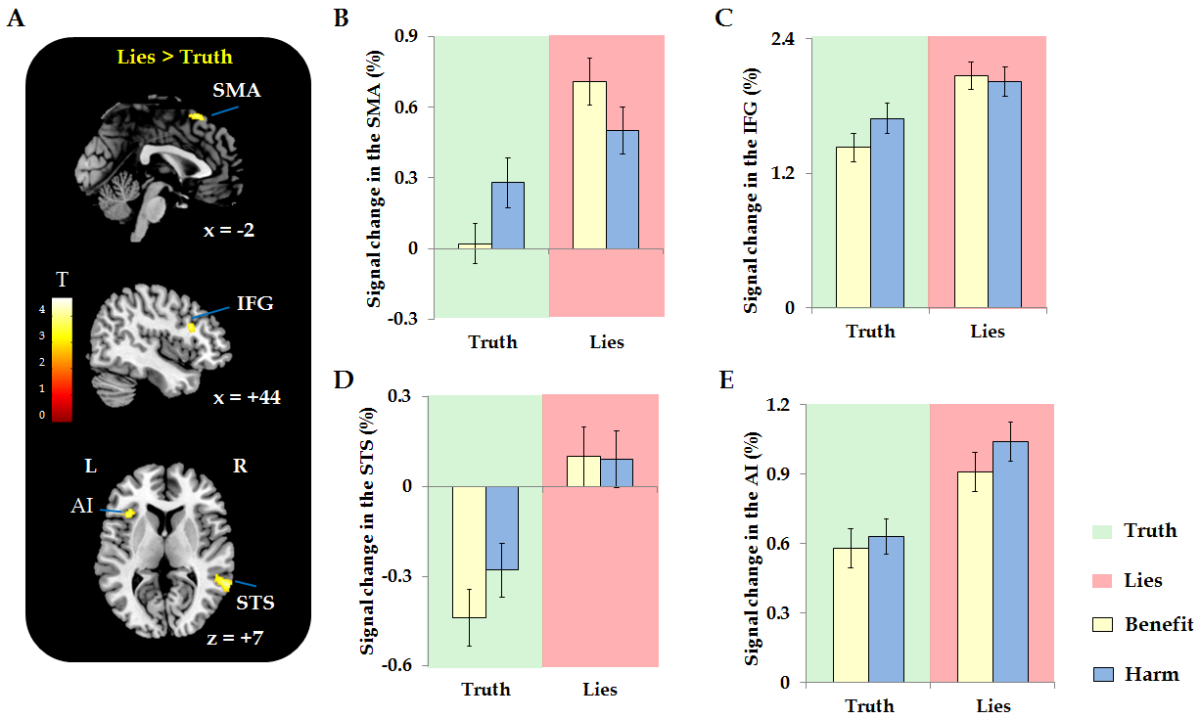

Figure 7.5: fMRI results in Study 3. (A) The brain regions were activated in the main effect of lies versus truth. Percentage signal changes were extracted from the left SMA (B), the right IFG (C), the right STS (D), and the left AI (E). (SMA: supplementary motor area; IFG: inferior frontal gyrus; STS: superior temporal sulcus; AI: anterior insula; error bars: s.e.m.)

7.4. Discussion

The aim of Study 3 was to investigate how altruistic outcomes influence the recipients' neural responses to lies and truth. The payoffs of lies and truth were manipulated to be more financially beneficial or harmful to participants to introduce altruistic outcomes and self-profit outcomes. Behaviorally, the recipients' emotional valence toward lies was more negative than truth. The beneficial outcomes reduced the difference in the valence between truth and lies. The behavioral results suggest that the outcomes modulate individuals' emotional responses to lies and truth.

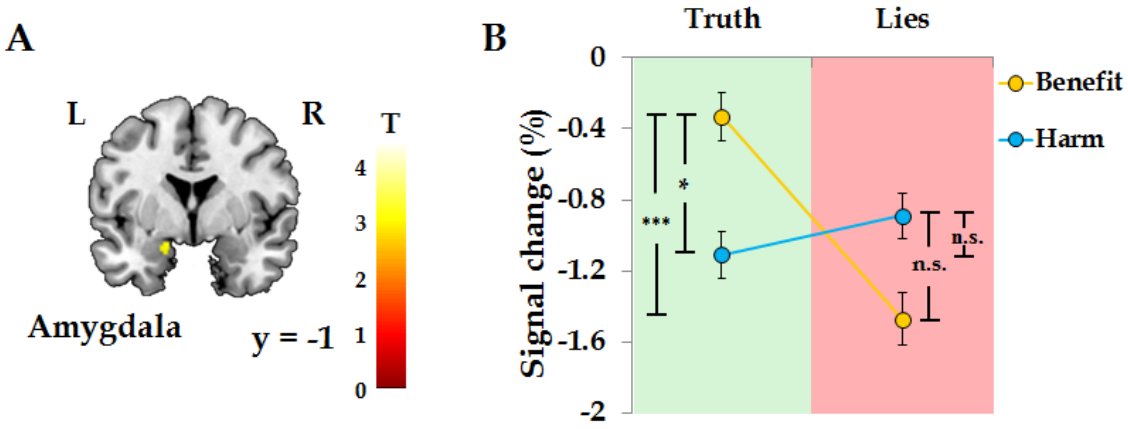

Figure 7.6: fMRI results in Study 3. (A) The left amygdala was activated in the interaction effect of (*beneficial truth - harmful truth*) versus (*beneficial lies - harmful lies*) after small-volume correction for multiple comparisons ($P_{FWE\text{-}corrected} < 0.05$). (B) Percentage signal changes were extracted from the left amygdala. Significant differences in brain activity were observed in the contrasts of *beneficial truth* versus *harmful truth* and *beneficial truth* versus *beneficial lies*. No significant differences were observed in the contrast of *beneficial lies* versus *harmful lies* and the contrast of *harmful lies* versus *harmful truth*. (* $P < 0.05$; *** $P < 0.001$; n.s.: not significant; error bars: s.e.m.)

At the neural level, the amygdala was activated in the interaction effect of (*beneficial truth - harmful truth*) versus (*beneficial lies - harmful lies*). Beneficial outcomes significantly increased amygdalar activity in the truth-telling condition. The amygdala plays an important role in the perception and production of emotion or affect (Öhman and Mineka, 2001; Phelps *et al.*, 2001; Mason *et al.*, 2006). Previous studies found that the amygdala is crucial in the cognitive and intentional control of mood (Brühl *et al.*, 2014; Young *et al.*, 2014). Amygdalar activity was increased in the successful regulation of negative affect (Diekhof *et al.*, 2011) and was decreased in the presence of reappraisal (Buhle *et al.*, 2014). In addition to the emotional responses and regulations, the amygdala is also associated with reward processing (Gaffan *et al.*, 1988; Cador *et al.*, 1989; Holland and Gallagher, 1999; Ernst *et al.*, 2005). The

emotional experiences of social outcomes (i.e., lies and truth with different monetary outcomes) might modulate activity in the amygdala.

In the main effect of lies versus truth, the anterior insula was activated. The insula is another important region in responding to negative events, and has been implicated in the experience of monetary loss (O'Doherty *et al.*, 2003), aversion (Paulus *et al.*, 2003; Chang *et al.*, 2011), disgust (Calder *et al.*, 2000; Sprengelmeyer, 2007), and unfairness (Sanfey *et al.*, 2003). The aversive feeling elicited by an immoral mean might elicit higher activity in the anterior insula. In Study 2, the anterior insula was found to be sensitive to the modulation of altruistic outcomes on moral emotions or moral evaluation of the behaviors. However, in Study 3, the anterior insula was not activated in the interaction effect of the outcome and mean. In the ratings of moral acceptance, lies were judged to be more morally unacceptable, regardless of the outcomes. That might explain why the anterior insula was only activated in the main effect rather than in the interaction effects, whereas the amygdala might be more sensitive to the basic emotional responses to different social outcomes.

The comparison of lying versus truth-telling also activated the STS, which is critical for social perception, i.e., the evaluation of the social communicative intentions of others extracted from gaze, facial expressions, body gestures, and motions (Pelphrey *et al.*, 2004; Moll *et al.*, 2005). In addition, the posterior STS is also critical for moral sensitivity (Robertson *et al.*, 2007). The posterior STS was more active when individuals were judging the facial trustworthiness (Bzdok *et al.*), viewing moral pictures (Harenski and Hamann, 2006), and making moral judgments (Moll *et al.*, 2002a; Greene *et al.*, 2004; Heekeren *et al.*, 2005; Sevinc and Spreng, 2014). In previous deception studies, STS activation was observed when participants judged the behaviors to be misleading (Grezes *et al.*, 2004; Grezes *et al.*, 2006). The recognition of morally transgressive behaviors (e.g., lying) might elicit higher activation in the STS (Robertson *et al.*, 2007).

The truth-telling behaviors and the beneficial outcomes were perceived as more positive and morally acceptable. The main effects of decision (i.e., truth > lies) and outcome (i.e., benefit > harm) activated the NAcc. The NAcc codes the processing of monetary rewards (Knutson and Cooper, 2005; Adcock *et al.*, 2006; Sabatinelli *et al.*, 2007). In addition to monetary rewards, the NAcc has been identified in the studies of social rewards (Izuma *et al.*, 2008; Spreckelmeyer *et al.*, 2009; Häusler *et al.*, 2015). In the prisoner's dilemma game, the faces of intentional cooperators elicited higher activity in the NAcc compared to the neutral faces (Singer *et al.*, 2004a). Therefore, in Study 3, the social (being honestly treated) and monetary rewards (beneficial outcomes) might elicit higher activity in the NAcc.

7.5. Limitations

In the beneficial conditions in Study 3, because the decisions made by the senders earned the receivers more money, the intention of senders might be perceived as altruistic. However, the senders' decisions not only profited the receivers but also the senders. The setting might weaken the effect of the benevolence in beneficial lies if recipients assumed that the senders lied to them out of consideration for their own interests rather than the recipients' interests.

IV. General discussion

In this dissertation, three research questions were investigated. Study 1 was used to investigate whether different neural correlates were involved in the (dis)honest decision-making process in the spontaneous and instructed paradigms. Study 2 was used to investigate the impact of altruistic or self-profit outcomes on the deceptive decision-making process and the underlying neural mechanisms from the perspective of the communicators. Study 3 was used to investigate the impact of beneficial and harmful outcomes on the neural processes underlying lies and truth from the perspective of the recipients.

In Section 1, the key findings from the three studies are first highlighted (Section 8.1), followed by a discussion of the differences between laboratory and real-life lies (Section 8.2), the cognitive demands of lies and truth in different contexts and different individuals (Section 8.3), the impact of altruistic outcomes on the process of deception (Section 8.4), and the directions for future studies of deception (Section 8.5). Finally, a summary of this dissertation is presented in Section 9.

8. Discussion

8.1. Overview of the key results from the three studies

In Study 1, the most important finding at the behavioral level is that the untruthful responses in the spontaneous paradigm (i.e., spontaneous lying/lies) were judged to be closer to the concept of lies than the untruthful responses in the instructed paradigm (i.e., instructed lying/lies). Regarding the emotional valence, making truthful responses was more positive than making untruthful responses in the spontaneous paradigm, whereas the opposite pattern was observed in the instructed paradigm. At the neural level, the right ventral lateral prefrontal cortex (VLPFC), the right ventral medial prefrontal cortex (VMPFC), and the right inferior parietal lobule (IPL) were significantly activated by spontaneous truthful responses in partially dishonest participants. Compared with the spontaneous truthful responses in honest participants, the VLPFLC, the VMPFC, and the IPL were also activated by spontaneous truthful responses in partially dishonest participants.

In Study 2, participants lied more often for the benefit of charitable organizations. At the neural level, activation was observed in the right anterior insula (AI) in the comparison of lying for oneself and lying for a charity in the conditions with honesty concerns to obtain high payoffs. Moreover, AI activity reflects individual differences in the altruistic impact on honesty concerns (to which extent the charity-profit condition reduced the effect of honesty concerns).

In Study 3, truth-telling and beneficial outcomes were judged to be more emotionally positive and more morally acceptable than lying. Beneficial outcomes decreased the valence differences between truth-telling and lies compared with harmful outcomes. At the neural level, the comparisons of truth-telling versus lies and beneficial outcomes versus harmful outcomes activated the nucleus accumbens (NAcc). In the comparison of lies and truth-telling, the left anterior insula was activated. A

significant interaction between the outcome and mean was observed in the left amygdala, further showing that beneficial truth-telling elicited higher activity in the left amygdala.

8.2. The differences between laboratory lies and real-life lies

When we consider whether it is feasible to apply the findings from the neuroimaging studies to detect lies in real life or even in the court, the weak reliability (i.e., the extent to which the studies produce stable and consistent results), external validity (i.e., the extent to which the results of the studies can be generalized to other settings, to other individuals, and over time), and construct validity (i.e., the extent to which the studies test what they were designed to test) of the laboratory studies were the focus of the debate (Schauer, 2010). Differences exist between lies in the laboratory and real-life lies, particularly when the instructed paradigms were used to investigate deception in the laboratory (Kanwisher, 2009; Phelps, 2009).

Many human behaviors and decisions are guided by the motivation to achieve pleasant states and avoid unpleasant states (Daw *et al.*, 2006; Linke *et al.*, 2010). The motives to lie can be defined according to three dimensions (Vrij, 2007): (1) for one's own benefit (self-oriented) or for the benefit of others (other-oriented), (2) to gain advantage or to avoid costs, and (3) for materialistic reasons or for psychological reasons. However, in most instructed studies, the motives to lie are relatively weak. It might be better to use mock-crime scenarios[25] to investigate lying behaviors because participants might be more involved in the experiment and the situation might be more similar to a real life situation. However, even in the mock-crime scenarios,

[25] In the task, participants are usually assigned to a mock-crime group or a no-crime group. Participants in the mock-crime group conduct a mock crime (e.g., gun shooting in the hospital (Mohamed *et al.*, 2006) or damaging and stealing CDs (Kozel *et al.*, 2009a)). Participants in the no-crime group do not conduct any mock crime. After that, participants answer questions about the details of the crime, personal information, and whether they committed the crime. Participants are instructed to make truthful or untruthful responses.

participants were well aware that the mock-crime situations were artificial (Ganis and Keenan, 2009). Previous studies found that the situation where participants had the opportunity to make a choice elicited individuals' affective responses, which were associated with increased activity in the corticostriatal region, compared to the situation where participants did not have the opportunity to make a choice (Leotti *et al.*, 2010; Leotti and Delgado, 2011). In the motivational states, choices were made by the interactions between regions in the corticostriatal network (the prefrontal cortex and the striatum), which was different from the states without motivation (Leotti *et al.*, 2010).

With the development of neuroeconomics [26] and experimental paradigms, spontaneous paradigms have been used more often in neuroimaging studies of deception. In contrast to the instructed paradigms, participants are allowed to freely decide to lie or to tell the truth. The representative studies are from Greene and Paxton (Greene and Paxton, 2009) and Baumgartner and his colleagues (Baumgartner *et al.*, 2009). In the spontaneous paradigms, participants are usually placed in the incentivizing contexts, where they are tempted to deceive others and earn more benefits for themselves. The mental processes of spontaneous lies, such as cognitive control (Greene and Paxton, 2009; Zhu *et al.*, 2014) and emotional responses (Ekman, 1985, 1989), might be different from the mental processes of instructed lies. A direct comparison of untruthful responses in these two types of experimental paradigms was performed in Study 1. The subjective judgments about two types of untruthful responses showed that spontaneous lies were closer to the concept of "lies" than instructed lies. The results suggest that the motivation to lie and spontaneous choices are essential to reduce the difference between lies in a laboratory environment and lies in our daily life.

[26] Neuroeconomics is an interdisciplinary field that aims to investigate human decision-making process and develop theories to understand human behaviors. It combines research methods from neuroscience, economics, and psychology.

8.3. Lying and truth-telling: which is more cognitively demanding?

In many previous functional neuroimaging studies of deception, many brain regions, particularly the executive brain regions, were activated when participants were giving untruthful responses. The DLPFC is one of the regions that was activated in the comparison of instructed lies and truth (Lee *et al.*, 2002; Lee *et al.*, 2005; Luan Phan *et al.*, 2005; Nunez *et al.*, 2005; Abe *et al.*, 2006; Abe *et al.*, 2007). The disruption of the right DLPFC decreased the dishonest response rate (Karton and Bachmann, 2011; Karton *et al.*, 2014), and facilitating the right DLPFC increased the dishonest response rate (Karton *et al.*, 2014). However, truthful responses rarely elicited higher neural activity than untruthful responses (Christ et al., 2009; Farah et al., 2014). Similar activation patterns have been observed not only in the studies that used the instructed paradigms but also in some studies of spontaneous lying and truth-telling (Bhatt *et al.*, 2010; Sip *et al.*, 2010). In a study of broken promises (Baumgartner *et al.*, 2009), the left DLPFC was activated when dishonest participants were making promises that they intended to break compared with the promises they were not going to break. The DLPFC exerts a key function in tasks that require a high level of cognitive control (Sanfey *et al.*, 2003; Aron *et al.*, 2004; Spitzer *et al.*, 2007). In studies of self-control, disruption of the right DLPFC weakened individuals' ability to resist economic temptation (Knoch *et al.*, 2006b) and increased individuals' risky behaviors (Knoch *et al.*, 2006a; Knoch and Fehr, 2007). The previous findings seem to support the notion that lying might be more cognitively demanding, whereas truth-telling is more like a default behavior (Spence et al., 2004), and the greater involvement of self-control during lying evokes higher DLPFC activity.

Section 2.1 introduced the findings from studies of the reaction times for lying and truth-telling. Compared with truth-telling, lying seems to require more time. However, practice can reduce the response times for lying (Walczyk *et al.*, 2009; Verschuere *et al.*, 2011; Hu *et al.*, 2012). In addition, dishonest behaviors were decreased in the situation where the response time was not restricted compared with the situation

where the response time was restricted (Shalvi *et al.*, 2012). Moreover, self-depletion can increase the frequencies of lying (Mead *et al.*, 2009; Gino *et al.*, 2011; Kouchaki and Smith, 2014). These findings suggest that the cognitive demands of dishonesty can be modulated and honesty also requires more time and cognitive control in certain contexts. In Study 1, no significant difference in the reaction times was observed between lying and truth-telling. In addition, greater involvement of the right VLPFC, the right DLPFC, and the right IPL was observed when partially dishonest participants spontaneously provided truthful responses that led to a monetary loss. In previous studies, the prefrontal cortex was also believed to participate in making honest decisions, particularly in the experiments that used the spontaneous paradigms. In a coin-flip task (Greene and Paxton, 2009; Abe and Greene, 2014), participants predicted the outcomes of coin flips. Correct predictions led to monetary benefits. In the condition without the opportunity to cheat, participants had to explicitly report their predictions and then indicated whether their predictions were correct. In the condition with the cheating opportunity, participants did not have to explicitly report their predictions, and, therefore, they could earn more money by lying about their predictions. When dishonest individuals were lying and when they were telling the truth, activation was observed in the control-related regions, including the DLPFC and the VLPFC. In one lesion study (Zhu *et al.*, 2014), individuals with lesions in the DLPFC showed reduced honesty concerns and lied more often when playing the sender-receiver game as senders compared with OFC lesioned patients and healthy controls. The researchers further speculated that the engagement of the DLPFC during lying might reflect the active but ultimately unsuccessful involvement of cognitive control. These findings suggest that the DLPFC plays a critical role in both the dishonest decision-making process and the honest decision-making process.

Compared with lying, telling the truth can be more cognitively demanding to some individuals in some circumstances. The cognitive resources required for lying and truth-telling might be context- and individual-dependent. In the contexts with a weak

motivation to lie (e.g., the lack of monetary or social rewards), truth-telling might be more like a default behavior and would, therefore, require fewer cognitive resources. However, in the contexts with a strong motivation to lie (e.g., avoiding punishments or gaining rewards), truth-telling might require additional cognitive control to fight against the temptation of dishonest gain. Moreover, different individuals might have different allocations of cognitive resources when they are making the decisions. In Study 1, partially dishonest participants required greater involvement of the VLPFC, the DLPFC, and the IPL than honest participants when they were telling the truth in the spontaneous session to forfeit monetary gains. Similar results were reported by Greene and Paxton (2009). Honest individuals might require less cognitive effort to make honest decisions compared with dishonest individuals who might expend more effort to forfeit immoral gains.

8.4. The impact of altruistic outcomes on the process of deception

Mounting evidence suggests that lying is psychologically costly (Lundquist *et al.*, 2009). Previous studies found that introducing honesty concerns in economic decision-making made people choose the materially advantageous allocation less often (Gneezy, 2005) and people tended to avoid the situations with honesty concerns (Shalvi *et al.*, 2011b). Even if the lies would not be punished and would not cause any financial harm to others, many individuals still forfeited larger benefits and behaved honestly (Erat and Gneezy, 2012; López-Pérez and Spiegelman, 2013), suggesting that there are psychological costs of lying.

Nevertheless, the attitudes toward lying (e.g., acceptability) might be linked to its good or bad outcomes (DePaulo, 2004). From the perspective of the recipients and third-party observers, altruistic lies that benefit others were judged to be morally appropriate (Hayashi *et al.*, 2014; Levine and Schweitzer, 2014). In addition, the altruistic outcomes of lies can increase trust in liars who told prosocial lies (Levine and Schweitzer, 2015). From the perspectives of the communicators, a certain number

of participants chose to lie to profit others, even at the expense of their own payoff (Erat and Gneezy, 2012). Lying behavior was observed even more frequently if it benefited charities (Lewis *et al.*, 2012).

Moral goals or intentions influence people's decisions or judgments by modulating the negative attitudes or emotional responses toward the immoral means. The findings from moral judgment studies suggest that emotions played an integral role in the judgments toward moral dilemmas (Greene et al., 2001). The avoid-related and approach-related social emotions (such as disgust, admiration, and compassion) and associated emotional networks (such as the anterior cingulate, the anterior insula, and the hypothalamus) are vital to moral evaluation (Funk and Gazzaniga, 2009). Previous studies in the field of social neuroscience implicated the social brain network including the brain regions of the insula/ACC, the amygdala, the temporoparietal junction, the DLPFC, the VMPFC, and the striatum (Glimcher and Fehr, 2013). The increased activity in the amygdala when encountering beneficial truth-telling might reflect the process of a rewarding situation. The amygdala is involved in the processes of stimuli with positive valence and reward (Baxter and Murray, 2002; Murray, 2007; Anders *et al.*, 2008; Ball *et al.*, 2009; Morrison and Salzman, 2010). In animal studies, selective amygdala lesions can affect the ability to associate stimuli with reward value (Gaffan *et al.*, 1988; Cador *et al.*, 1989; Holland and Gallagher, 1999). In human studies, activity in the amygdala and the NAcc was increased for both adolescents and adults when they were winning money (as opposed to not winning money) (Ernst *et al.*, 2005).

Nevertheless, the amygdala also plays a major role in the perception and production of negative emotion or affect (Öhman and Mineka, 2001; Phelps *et al.*, 2001; Mason *et al.*, 2006). When perceiving stimuli of aversive content, healthy people show enhanced attention (compared with the perception of neutral content). However, lesions in the bilateral amygdala damage this improved attention toward emotionally significant stimuli (Anderson and Phelps, 2001), suggesting that the amygdala ensures that the emotional events capture enough attention. In the study of loss

aversion (i.e., the avoidance of choices that might lead to losses), people with lesions in the amygdala showed a reduction in loss aversion in a gambling task compared with the normal controls (De Martino *et al.*, 2010). In previous neuroimaging studies addressing the judgment of lying and truth-telling behaviors, the amygdala was more active when participants thought they were being deceived by the experimenter (Grezes *et al.*, 2006) and the amygdala was more active toward participants' own moral violation (Berthoz *et al.*, 2006). The role of the amygdala in the interaction between the mean and outcome might be associated with the process of the emotional responses to social outcomes.

Concerning the insula, it involves in multiple domains, such as perceptual decision-making (Binder *et al.*, 2004; Thielscher and Pessoa, 2007), deceptive decision-making (Christ *et al.*, 2009; Farah *et al.*, 2014; Lisofsky *et al.*, 2014), and empathy (Singer *et al.*, 2004b; Jackson *et al.*, 2005). In a broken promises study (Baumgartner *et al.*, 2009), the researchers used a trust game where participants (as trustees) could either make a promise about paying back money or have no chance to make any promise. They then waited for an investor's decision to trust them or not. If the investor trusted them, participants decided to pay back or keep the money. In the dishonest participants, promises of paying back the money (versus the no promise condition) elicited increased activity in the right anterior insula and the IFG during the promise stage, whereas the no promise condition elicited higher activity in the anterior insula and the IFG during the anticipation stage. The authors speculated that the former was associated with aversive emotional experiences of providing misleading promises, and the latter was due to the anticipation of negative and unforeseeable emotional events.

In the social domain, the anterior insula can detect deviations from the socially acceptable outcome (Glimcher and Fehr, 2013) and is associated with the prediction and detection of other aversive stimuli (Nitschke *et al.*, 2006; Caria *et al.*, 2010; Chang *et al.*, 2011; Liljeholm *et al.*, 2014). In the ultimatum game, the introduction of sad emotions (Harlé *et al.*, 2012) and the reappraisal of the proposer's intentions as

more negative (Grecucci *et al.*, 2012) increased the responders' rejection rates toward unfair offers and activity in the anterior insula when facing unfair offers. The tendency to accept unfair proposals was associated with decreased activity in the anterior insula, which has been implicated in negative affect (Tabibnia *et al.*, 2008).

According to the somatic markers' hypothesis, the bodily emotional systems react to a certain act before the cognitive process reacts to its outcomes (Damasio, 1994). The anterior insula is one of the key brain regions representing the somatic state that arises in emotion processing, particularly negative emotions (Bechara and Damasio, 2005; Naqvi *et al.*, 2006), and provides knowledge for making fast and advantageous economic decisions (Bechara and Damasio, 2005). When individuals encounter moral events, the anterior insula might detect the transgressions in the events by generating negative or aversive feelings.

8.5. Directions for future studies

Studies of deception in the laboratory settings are informative and improve our understanding of deception in the real world. The fMRI-based lie detection method is applied according to the findings from the neuroimaging studies of lying. In some cases, lying elicits higher activation in multiple brain regions compared with truth-telling. However, in the context where truth-telling is more cognitively demanding, honest responses might be mistakenly classified as lies (Farah *et al.*, 2014). When an individual is responding to a question, a lack of activation in areas associated with inhibition or cognitive control cannot confirm that the individual is telling the truth (Meijer *et al.*, 2016). Therefore, the paradigms used in the laboratory experiments to examine deception are important (Ganis and Keenan, 2009), and the gaps between the laboratory lies and lies in real life should be reduced.

Several issues should be considered when researchers are conducting experiments to investigate the neural correlates of deception. First, whether (dis)honest decisions are

made on participants' own initiative should be taken into consideration. The neural processes of instructed lies differ from the neural processes of spontaneous lies. Lies in the experimental settings where participants can spontaneously make their decisions to lie, compared with lies in the settings where participants can only follow others' instructions to lie, might be more similar to the lies in real life. Spontaneous paradigms provide us knowledge about lies in real life. In some spontaneous paradigms, to identify when participants lie in the scanner, participants' behaviors are under observation. However, participants might make different decisions if they know that they are under observation. To reduce the effects caused by the observation, experimenters can tell participants that experimenters are not observing while participants are making decisions and their data will be analyzed by researchers who have no access to their personal information. Nevertheless, it is important to develop new paradigms in which participants can privately make decisions and participants' behaviors can be monitored during the experiment.

Second, manipulations, such as inducing social interaction and punishment for dishonesty, should be considered. Social interaction involved in the experiment is important for understanding lying in the context where participants interact with others. In many instructed studies, participants lied to a investigator/experimenter (Davatzikos *et al.*, 2005; Langleben *et al.*, 2005; Lee *et al.*, 2005; Luan Phan *et al.*, 2005), or participants did not lie to a specific person (Ganis *et al.*, 2003; Lee *et al.*, 2010). The context with social interaction might have different impacts on individuals with different social preferences and make individuals take the perspective of opponents, and, therefore, social interaction influence the neural correlates of lying (Lisofsky *et al.*, 2014). By adopting interactive games (Sip *et al.*, 2010; Volz *et al.*, 2015), cognitive processes involved in social interaction during lying can be further investigated. Moreover, punishment for dishonesty is an important manipulated variable. Punishment for dishonesty might influence the decision-making process of lying through modulating the expected value of dishonest behaviors.

Third, different kinds of lies should be investigated, and some other techniques should be used to investigate the neural mechanisms underlying deception. Different motivations and situations influence the psychological and neural mechanisms underlying lying and truth-telling. Monetary or social outcomes caused by lies can be manipulated to investigated different kinds of lies. Moreover, techniques like the skin conductance response, transcranial direct current stimulation (tDCS), and transcranial magnetic stimulation (TMS) provide us information about participants' physiologically arousing states and the effects caused by disruption or facilitation of a specific brain region.

9. Summary

Deception is an important social phenomenon. A paradigm of "instructed lies" was used in many previous functional magnetic resonance imaging (fMRI) studies of deception. Instructed lies are different from other, more spontaneous lying behaviors. Study 1 investigated whether there are different neural mechanisms underlying the deceptive decision-making process in the spontaneous and the instructed paradigms. A modified *sic bo* gambling game was used to incentivize lying and enhance participants' involvement, particularly in the spontaneous paradigm. In the spontaneous session, participants freely decided whether to lie, whereas in the instructed session they were explicitly instructed to make either truthful or untruthful responses. Spontaneous untruthful responses were closer to the concept of "lies" than instructed untruthful responses. In addition, at the neural level, the right dorsal lateral prefrontal cortex, the right ventral lateral prefrontal cortex, and the right inferior parietal lobule were more strongly activated by spontaneous truth-telling (versus other decisions, including spontaneous lying, instructed lying, and instructed truth-telling). These regions were also activated when comparing the spontaneous truth told by dishonest participants with the spontaneous truth told by honest participants. The results of Study 1 suggest that the extra cognitive control required to suppress the motives of benefiting oneself by lying evokes greater involvement of the frontoparietal network.

Study 2 and Study 3 were performed to investigate the neural correlates of the impact of altruistic outcomes on the process of deception from the perspectives of communicators and recipients, respectively. In Study 2, a modified sender-receiver paradigm was adopted, where participants could earn different amounts of money by lying or telling the truth to benefit either a charity or themselves. Participants lied more often to benefit charities than to benefit themselves. The altruistic impact on honesty concerns is associated with the anterior insula, which was activated when participants lied to benefit themselves (versus charitable organizations). Furthermore,

anterior insula activity reflected individual differences in the modulation of the altruistic outcomes on honesty concerns. The results suggest that the anterior insula is the key hub for integrating information from different streams of outcomes and acts to detect the deviations in socially or morally acceptable acts.

Study 3 was performed to investigate how monetary outcomes influence the neural responses to lies and truth. A modified sender-receiver game was used, where participants were the direct recipients of lies and truthful statements with beneficial or harmful monetary outcomes. Both truth-telling and beneficial outcomes elicited higher activity in the nucleus accumbens, suggesting that the nucleus accumbens is associated with the process of monetary and social rewards. Compared with truth, lies elicited higher activity in the supplementary motor area, the inferior frontal gyrus, the superior temporal sulcus, and the anterior insula. Moreover, the beneficial monetary outcomes enhanced the neural activity in the amygdala in the truth-telling condition. The results identified a neural network associated with the reception of lies and truth with different outcomes, including the regions linked to the reward process, recognition, and emotional experiences of being treated (dis)honestly.

V. Abstract

10. English abstract

Investigating the process of deception is crucial for our understanding of lying behaviors. In this dissertation, three studies were performed to investigate: 1) the neural bases of lying and truth-telling in two different experimental paradigms and 2) the impact of the altruistic outcomes (i.e., the outcomes of the acts that financially benefit others) on the processes of lies and truth. In Study 1, participants provided (un)truthful responses either on one's own initiative in the spontaneous paradigm or by following others' instructions in the instructed paradigm. The behavioral results suggest that the free choice of making one's own decisions is one of the key components of the concept of "lies." At the neural level, the ventral lateral prefrontal cortex, the dorsal lateral prefrontal cortex, and the inferior parietal lobe showed different activation patterns in the two different paradigms. The results suggest that these regions might provide cognitive control over the temptation of dishonest gain, particularly in the paradigms that allow individuals to freely make their decisions. In Study 2 and Study 3, the outcomes of lying/truth-telling behaviors were manipulated to investigate the neural correlates of the impact of altruistic outcomes on the processes of the behaviors in both the communicators and the recipients. The results showed that the altruistic outcomes of moral behaviors mainly modulated neural activity in the nucleus accumbens, the amygdala, and the anterior insula. The nucleus accumbens might be sensitive to both social rewards and monetary rewards. The amygdala might be involved in generating emotional responses to social outcomes, whereas the anterior insula might code deviations from socially or morally acceptable acts. Taken together, the results suggest that the neural processes underlying deception in the frontoparietal network, the limbic system, the mesolimbic system, and the insula cortex are associated with the psychological processes of deception, including cognitive control, reward coding, and emotional responses. The findings extend our knowledge of the neural processes underlying lies and truth in different contexts and with different outcomes.

11. German abstract

Die Untersuchung von Täuschungsprozessen ist für das Verständnis des Themengebiets des Lügens unabdingbar. Im Rahmen dieser Dissertation wurden drei Studien durchgeführt, um 1) die neuronale Basis von Lügen und wahrheitsgemäßen Angaben mithilfe von zwei unterschiedlichen experimentellen Paradigmen zu untersuchen, sowie 2) den Einfluss von altruistischen Ergebnissen (d.h. die Ergebnisse von Entscheidungen, die Anderen einen finanziellen Vorteil verschaffen) auf die Prozesse von „Lügen“ und „Wahrheit sagen“ zu eruieren. In Studie 1 konnten die Probanden in dem „spontanen Paradigma“ selbst bestimmen, ob sie wahre oder unwahre Angaben machen, während sie in dem „instruierten Paradigma“ von Anderen die Vorgabe für diese Entscheidung erhielten. Die behavioralen Ergebnisse von Studie 1 legen nahe, dass die freie Wahl der eigenen Entscheidungen eine der Schlüsselkomponenten des Konzepts des „Lügens“ darstellt. Auf der neuronalen Ebene zeigten der ventrolaterale präfrontale Kortex, der dorsolaterale präfrontale Kortex und der untere Parietallappen unterschiedliche Aktivierungsmuster in den beiden Paradigmen. Die Ergebnisse deuten darauf hin, dass diese Regionen kognitive Kontrolle über die Versuchung von unehrlichem Gewinn bieten können, insbesondere in den Paradigmen, die es Individuen erlauben, freie Entscheidungen zu treffen. In den Studien 2 und 3 wurden die aus betrügerischem/wahrheitsgemäßem Verhalten resultierenden Ergebnisse manipuliert, um die neuronalen Korrelate des Einflusses von altruistischen Ergebnissen auf das Verhalten des möglichen Lügners und des Empfängers zu untersuchen. Die Resultate zeigen, dass altruistische Motive des moralischen Verhaltens hauptsächlich die neuronale Aktivität im Nucleus accumbens, der Amygdala und der anterioren Insula modulierten. Der Nucleus accumbens könnte sowohl auf soziale Belohnungen (ehrlich behandelt zu werden) als auch auf monetäre Belohnungen (höhere monetäre Zahlungen zu erhalten) sensibel reagieren. Die Amygdala könnte im Erzeugen emotionaler Antworten auf soziale Ergebnisse involviert sein, wohingegen die anteriore Insula Abweichungen von sozial oder

moralisch akzeptablen Handlungen kodieren könnte. Zusammengenommen legen die Ergebnisse nahe, dass die neuronalen Täuschungsprozesse im frontoparietalen Netzwerk, im limbschen System, im mesolimbischen System und der Insula mit den psychologischen Prozessen der Täuschung, inklusive kognitiver Kontrolle, Belohnungskodierung und emotionaler Antworten assoziiert sind. Die Befunde erweitern unser Verständnis über die zu Grunde liegenden Prozesse von Lügen und Wahrhaftigkeit in unterschiedlichen Kontexten und unterschiedlichen Zielen.

References

Abe, N., Fujii, T., Ito, A., Ueno, A., Koseki, Y., Hashimoto, R., Hayashi, A., Mugikura, S., Takahashi, S., & Mori, E. (2014) The neural basis of dishonest decisions that serve to harm or help the target. Brain and Cognition, 90(0), 41-49.

Abe, N., & Greene, J. D. (2014) Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. Journal of Neuroscience, 34(32), 10564-10572.

Abe, N., Okuda, J., Suzuki, M., Sasaki, H., Matsuda, T., Mori, E., Tsukada, M., & Fujii, T. (2008) Neural correlates of true memory, false memory, and deception. Cerebral Cortex, 18(12), 2811-2819.

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(2), 287-295.

Abe, N., Suzuki, M., Tsukiura, T., Mori, E., Yamaguchi, K., Itoh, M., & Fujii, T. (2006) Dissociable roles of prefrontal and anterior cingulate cortices in deception. Cerebral Cortex, 16(2), 192-199.

Abeler, J., Becker, A., & Falk, A. (2014) Representative evidence on lying costs. Journal of Public Economics, 113, 96-104.

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. (2006) Reward-motivated learning: mesolimbic activation precedes memory formation. Neuron, 50(3), 507-517.

Alexander, A. (2007) Functional Magnetic Resonance Imaging Lie Detection: Is a Brainstorm Heading toward the Gatekeeper. Hous. J. Health L. & Pol'y, 7, 1-56.

Amaro, E., & Barker, G. (2006) Study design in fMRI: Basic principles. Brain and Cognition, 60(3), 220-232.

Anders, S., Eippert, F., Weiskopf, N., & Veit, R. (2008) The human amygdala is sensitive to the valence of pictures and sounds irrespective of arousal: an fMRI study. Social Cognitive and Affective Neuroscience, 3(3), 233-243.

Anderson, A. K., & Phelps, E. A. (2001) Lesions of the human amygdala impair enhanced perception of emotionally salient events. Nature, 411(6835), 305-309.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004) Inhibition and the right inferior frontal cortex. Trends in Cognitive Sciences, 8(4), 170-177.

Ball, T., Derix, J., Wentlandt, J., Wieckhorst, B., Speck, O., Schulze-Bonhage, A., & Mutschler, I. (2009) Anatomical specificity of functional amygdala imaging of

responses to stimuli with positive and negative emotional valence. Journal of Neuroscience Methods, 180(1), 57-70.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013) The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. NeuroImage, 76, 412-427.

Batson, C. D., & Shaw, L. L. (1991) Evidence for altruism: Toward a pluralism of prosocial motives. Psychological inquiry, 2(2), 107-122.

Battigalli, P., Charness, G., & Dufwenberg, M. (2013) Deception: The role of guilt. Journal of Economic Behavior & Organization, 93, 227-232.

Baumeister, R. F., Vohs, K. D., DeWall, C. N., & Zhang, L. (2007) How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. Personality and Social Psychology Review, 11(2), 167-203.

Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., & Fehr, E. (2009) The neural circuitry of a broken promise. Neuron, 64(5), 756-770.

Baumgartner, T., Gianotti, L. R., & Knoch, D. (2013) Who is honest and why: Baseline activation in anterior insula predicts inter-individual differences in deceptive behavior. Biological Psychology, 94(1), 192-197.

Baxter, M. G., & Murray, E. A. (2002) The amygdala and reward. Nature Reviews Neuroscience, 3(7), 563-573.

Bechara, A., & Damasio, A. R. (2005) The somatic marker hypothesis: A neural theory of economic decision. Games and Economic Behavior, 52(2), 336-372.

Becker, G. S. (1968) Crime and Punishment: An Economic Approach. The Journal of Political Economy, 76(2), 169-217.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001) Predictability modulates human brain response to reward. The Journal of Neuroscience, 21(8), 2793-2798.

Berthoz, S., Grezes, J., Armony, J., Passingham, R., & Dolan, R. (2006) Affective response to one's own moral violations. NeuroImage, 31(2), 945-950.

Bhatt, M. A., Lohrenz, T., Camerer, C. F., & Montague, P. R. (2010) Neural signatures of strategic types in a two-person bargaining game. Proceedings of the National Academy of Sciences, 107(46), 19720-19725.

Bhatt, S., Mbwana, J., Adeyemo, A., Sawyer, A., Hailu, A., & Vanmeter, J. (2009) Lying about facial recognition: an fMRI study. Brain and Cognition, 69(2), 382-390.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004) Neural correlates of sensory and decision processes in auditory object identification. Nature Neuroscience, 7(3), 295-301.

Browndyke, J. N., Paskavitz, J., Sweet, L. H., Cohen, R. A., Tucker, K. A., Welsh-Bohmer, K. A., Burke, J. R., & Schmechel, D. E. (2008) Neuroanatomical correlates of malingered memory impairment: event-related fMRI of deception on a recognition memory task. Brain Injury, 22(6), 481-489.

Brühl, A. B., Scherpiet, S., Sulzer, J., Stämpfli, P., Seifritz, E., & Herwig, U. (2014) Real-time neurofeedback using functional MRI could improve down-regulation of amygdala activity during emotional stimulation: a proof-of-concept study. Brain Topography, 27(1), 138-148.

Buhle, J. T., Silvers, J. A., Wager, T. D., Lopez, R., Onyemekwu, C., Kober, H., Weber, J., & Ochsner, K. N. (2014) Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. Cerebral Cortex, 24(11), 2981-2990.

Bzdok, D., Langner, R., Caspers, S., Kurth, F., Habel, U., Zilles, K., Laird, A., & Eickhoff, S. B. (2011) ALE meta-analysis on facial judgments of trustworthiness and attractiveness. Brain Structure and Function, 215(3-4), 209-223.

Cador, M., Robbins, T., & Everitt, B. (1989) Involvement of the amygdala in stimulus-reward associations: interaction with the ventral striatum. Neuroscience, 30(1), 77-86.

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000) Impaired recognition and experience of disgust following brain injury. Nature Neuroscience, 3(11), 1077-1078.

Caria, A., Sitaram, R., Veit, R., Begliomini, C., & Birbaumer, N. (2010) Volitional control of anterior insula activity modulates the response to aversive stimuli. A real-time functional magnetic resonance imaging study. Biological Psychiatry, 68(5), 425-432.

Carson, T. L. (2010) *Lying and deception: theory and practice*, Oxford University Press.

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011) Triangulating the neural, psychological, and economic bases of guilt aversion. Neuron, 70(3), 560-572.

Chisholm, R. M., & Feehan, T. D. (1977) The intent to deceive. The Journal of Philosophy, 74(3), 143-159.

Christ, S., Van Essen, D., Watson, J., Brubaker, L., & McDermott, K. (2009) The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. Cerebral Cortex, 19(7), 1557-1566.

Coleman, L., & Kay, P. (1981) Prototype semantics: The English word lie. Language, 57(1), 26-44.

Cui, Q., Vanman, E. J., Wei, D., Yang, W., Jia, L., & Zhang, Q. (2014) Detection of deception based on fMRI activation patterns underlying the production of a

deceptive response and receiving feedback about the success of the deception after a mock murder crime. Social Cognitive and Affective Neuroscience, 9(10), 1472-1480.

Damasio, A. (1994) Descartes' error: Emotion, reason and the human brain. Putnam, New York, 195-201.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M., Loughead, J., Gur, R., & Langleben, D. (2005) Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. NeuroImage, 28(3), 663-668.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006) Cortical substrates for exploratory decisions in humans. Nature, 441(7095), 876-879.

De Martino, B., Camerer, C. F., & Adolphs, R. (2010) Amygdala damage eliminates monetary loss aversion. Proceedings of the National Academy of Sciences, 107(8), 3788-3792.

Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D., & Fiez, J. A. (2000) Tracking the hemodynamic responses to reward and punishment in the striatum. Journal of Neurophysiology, 84(6), 3072-3077.

DePaulo, B. M. (2004) The many faces of lies. The social psychology of good and evil, 303-326.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996) Lying in everyday life. Journal of Personality and Social Psychology, 70(5), 979-995.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003) Cues to deception. Psychological bulletin, 129(1), 74-118.

Diekhof, E. K., Geier, K., Falkai, P., & Gruber, O. (2011) Fear is only as deep as the mind allows: a coordinate-based meta-analysis of neuroimaging studies on the regulation of negative affect. NeuroImage, 58(1), 275-285.

Ding, X. P., Gao, X., Fu, G., & Lee, K. (2013) Neural correlates of spontaneous deception: A functional near-infrared spectroscopy (fNIRS) study. Neuropsychologia, 51(4), 704-712.

Eadington, W. R. (1999) The economics of casino gambling. The Journal of Economic Perspectives, 13(3), 173-192.

Eisenberg, N. (2000) Emotion, regulation, and moral development. Annual Review of Psychology, 51(1), 665-697.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences, 113, 7900-7905.

Ekman, P. (1985) *Telling lies: Clues to deceit in the marketplace, marriage, and politics*, New York, W. W. Norton.

Ekman, P. (1989) 'Why lies fail and what behaviors betray a lie', in J. C. Yuille (ed), *Credibility assessment*, Dordrecht, the Netherlands, Kluwer.

Ellingsen, T., Johannesson, M., Tjøtta, S., & Torsvik, G. (2010) Testing guilt aversion. Games and Economic Behavior, 68(1), 95-107.

Erat, S., & Gneezy, U. (2012) White lies. Management Science, 58(4), 723-733.

Ernst, M., Nelson, E. E., Jazbec, S., McClure, E. B., Monk, C. S., Leibenluft, E., Blair, J., & Pine, D. S. (2005) Amygdala and nucleus accumbens in responses to receipt and omission of gains in adults and adolescents. NeuroImage, 25(4), 1279-1291.

Farah, M. J., Hutchinson, J. B., Phelps, E. A., & Wagner, A. D. (2014) Functional MRI-based lie detection: scientific and societal challenges. Nature Reviews Neuroscience, 15(2), 123-131.

Fehr, E. (2009) Social preferences and the brain. Neuroeconomics: Decision making and the brain, 215-232.

Feldman, R. S., Forrest, J. A., & Happ, B. R. (2002) Self-presentation and verbal deception: Do self-presenters lie more? Basic and Applied Social Psychology, 24(2), 163-170.

Fischbacher, U., & Föllmi-Heusi, F. (2013) Lies in Disguise-an Experimental Study on Cheating. Journal of the European Economic Association, 11(3), 525-547.

Fu, G., Heyman, G. D., Chen, G., Liu, P., & Lee, K. (2015) Children trust people who lie to benefit others. Journal of Experimental Child Psychology, 129, 127-139.

Funk, C. M., & Gazzaniga, M. S. (2009) The functional brain architecture of human morality. Current Opinion in Neurobiology, 19(6), 678-681.

Gaffan, E., Gaffan, D., & Harrison, S. (1988) Disconnection of the amygdala from visual association cortex impairs visual reward-association learning in monkeys. The Journal of Neuroscience, 8(9), 3144-3150.

Ganis, G., & Keenan, J. P. (2009) The cognitive neuroscience of deception. Social Neuroscience, 4(6), 465-472.

Ganis, G., Kosslyn, S., Stose, S., Thompson, W., & Yurgelun-Todd, D. (2003) Neural correlates of different types of deception: an fMRI investigation. Cerebral Cortex, 13(8), 830-836.

Ganis, G., Morris, R. R., & Kosslyn, S. M. (2009) Neural processes underlying self-and other-related lies: an individual difference approach using fMRI. Social Neuroscience, 4(6), 539-553.

Ganis, G., Rosenfeld, J. P., Meixner, J., Kievit, R. A., & Schendan, H. E. (2011) Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging. NeuroImage, 55(1), 312-319.

Gaspar, J. P., & Schweitzer, M. E. (2013) The emotion deception model: a review of deception in negotiation and the role of emotion in deception. Negotiation and Conflict Management Research, 6(3), 160-179.

Gino, F., Ayal, S., & Ariely, D. (2013) Self-serving altruism? The lure of unethical actions that benefit others. Journal of Economic Behavior & Organization, 93, 285-292.

Gino, F., Schweitzer, M. E., Mead, N. L., & Ariely, D. (2011) Unable to resist temptation: How self-control depletion promotes unethical behavior. Organizational Behavior and Human Decision Processes, 115(2), 191-203.

Glimcher, P. W., & Fehr, E. (2013) *Neuroeconomics: Decision Making and the Brain*, New York, NY: Academic Press.

Gneezy, U. (2005) Deception: The role of consequences. The American Economic Review, 95(1), 384-394.

Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013) Measuring lying aversion. Journal of Economic Behavior & Organization, 93, 293-300.

Grecucci, A., Giorgetta, C., van't Wout, M., Bonini, N., & Sanfey, A. G. (2012) Reappraising the ultimatum: an fMRI study of emotion regulation and decision making. Cerebral Cortex, 23(2), 399-410.

Greely, H. T., & Illes, J. (2007) Neuroscience-based lie detection: The urgent need for regulation. American journal of law & medicine, 33(2-3), 377-431.

Greene, J., & Paxton, J. (2009) Patterns of neural activity associated with honest and dishonest moral decisions. Proceedings of the National Academy of Sciences, 106(30), 12506–12511.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004) The neural bases of cognitive conflict and control in moral judgment. Neuron, 44(2), 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001) An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105-2108.

Greening, S., Norton, L., Virani, K., Ty, A., Mitchell, D., & Finger, E. (2014) Individual differences in the anterior insula are associated with the likelihood of financially helping versus harming others. Cognitive, Affective, & Behavioral Neuroscience, 14(1), 266-277.

Grezes, J., Berthoz, S., & Passingham, R. (2006) Amygdala activation when one is the target of deceit: Did he lie to you or to someone else? NeuroImage, 30(2), 601-608.

Grezes, J., Frith, C., & Passingham, R. (2004) Inferring false beliefs from the actions of oneself and others: an fMRI study. NeuroImage, 21(2), 744-750.

Grèzes, J., Frith, C., & Passingham, R. E. (2004) Brain mechanisms for inferring deceit in the actions of others. The Journal of Neuroscience, 24(24), 5500-5505.

Hamann, S. (2012) Mapping discrete and dimensional emotions onto the brain: controversies and consensus. Trends in Cognitive Sciences, 16(9), 458-466.

Hare, T. A., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P., & Rangel, A. (2010) Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. The Journal of Neuroscience, 30(2), 583-590.

Harenski, C. L., & Hamann, S. (2006) Neural correlates of regulating negative emotions related to moral violations. NeuroImage, 30(1), 313-324.

Harlé, K. M., Chang, L. J., van't Wout, M., & Sanfey, A. G. (2012) The neural mechanisms of affect infusion in social economic decision-making: a mediating role of the anterior insula. NeuroImage, 61(1), 32-40.

Häusler, A. N., Becker, B., Bartling, M., & Weber, B. (2015) Goal or Gold: Overlapping Reward Processes in Soccer Players upon Scoring and Winning Money. PloS one, 10(4), e0122798.

Hayashi, A., Abe, N., Fujii, T., Ito, A., Ueno, A., Koseki, Y., Mugikura, S., Takahashi, S., & Mori, E. (2014) Dissociable neural systems for moral judgment of anti-and pro-social lying. Brain Research, 1556, 46-56.

Hayashi, A., Abe, N., Ueno, A., Shigemune, Y., Mori, E., Tashiro, M., & Fujii, T. (2010) Neural correlates of forgiveness for moral transgressions involving deception. Brain Research, 1332, 90-99.

Heekeren, H. R., Wartenburger, I., Schmidt, H., Prehn, K., Schwintowski, H. P., & Villringer, A. (2005) Influence of bodily harm on neural correlates of semantic and moral decision-making. NeuroImage, 24(3), 887-897.

Holland, P. C., & Gallagher, M. (1999) Amygdala circuitry in attentional and representational processes. Trends in Cognitive Sciences, 3(2), 65-73.

Hsu, M., Anen, C., & Quartz, S. R. (2008) The right and the good: distributive justice and neural encoding of equity and efficiency. Science, 320(5879), 1092-1095.

Hu, X., Chen, H., & Fu, G. (2012) A Repeated Lie Becomes a Truth? The Effect of Intentional Control and Training on Deception. Frontiers in psychology, 3.

Huettel, S. A., Song, A. W., & McCarthy, G. (2004) *Functional magnetic resonance imaging*, Sinauer Associates Sunderland.

Irtel, H. (2008) The PXLab Self-Assessment-Manikin Scales.

Ito, A., Abe, N., Fujii, T., Ueno, A., Koseki, Y., Hashimoto, R., Mugikura, S., Takahashi, S., & Mori, E. (2011) The role of the dorsolateral prefrontal cortex in deception when remembering neutral and emotional events. Neuroscience Research, 69(2), 121-128.

Izuma, K., Saito, D. N., & Sadato, N. (2008) Processing of social and monetary rewards in the human striatum. Neuron, 58(2), 284-294.

Jackson, P. L., Meltzoff, A. N., & Decety, J. (2005) How do we perceive the pain of others? A window into the neural processes involved in empathy. NeuroImage, 24(3), 771-779.

Jones, R. A., Schirmer, T., Lipinski, B., Elbel, G. K., & Auer, D. P. (1998) Signal undershoots following visual stimulation: A comparison of gradient and spin - echo BOLD sequences. Magnetic Resonance in Medicine, 40(1), 112-118.

Kahneman, D., Knetsch, J. L., & Thaler, R. (1986) Fairness as a constraint on profit seeking: Entitlements in the market. The American Economic Review, 728-741.

Kant, I. (1797) *On a supposed right to lie from philanthropy*, Cambrige: Cambridge University.

Kanwisher, N. (2009) The use of fMRI in lie detection: what has been shown and what has not. Using Imaging to Identify Deceit: Scientific and Ethical Questions, 7-13.

Karton, I., & Bachmann, T. (2011) Effect of prefrontal transcranial magnetic stimulation on spontaneous truth-telling. Behavioural Brain Research, 225(1), 209-214.

Karton, I., Rinne, J.-M., & Bachmann, T. (2014) Facilitating the right but not left DLPFC by TMS decreases truthfulness of object-naming responses. Behavioural Brain Research, 271, 89-93.

Knoch, D., & Fehr, E. (2007) Resisting the power of temptations. Annals of the New York Academy of Sciences, 1104(1), 123-134.

Knoch, D., Gianotti, L. R., Pascual-Leone, A., Treyer, V., Regard, M., Hohmann, M., & Brugger, P. (2006a) Disruption of right prefrontal cortex by low-frequency repetitive transcranial magnetic stimulation induces risk-taking behavior. Journal of Neuroscience, 26(24), 6469-6472.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006b) Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science, 314(5800), 829-832.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001a) Anticipation of increasing monetary reward selectively recruits nucleus accumbens. The Journal of Neuroscience, 21(16), 1-5.

Knutson, B., & Cooper, J. C. (2005) Functional magnetic resonance imaging of reward prediction. Current Opinion in Neurology, 18(4), 411-417.

Knutson, B., Fong, G. W., Adams, C. M., Varner, J. L., & Hommer, D. (2001b) Dissociation of reward anticipation and outcome with event-related fMRI. Neuroreport, 12(17), 3683-3687.

Knutson, B., Katovich, K., & Suri, G. (2014) Inferring affect from fMRI data. Trends in Cognitive Sciences, 18(8), 422-428.

Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000) FMRI visualization of brain activity during a monetary incentive delay task. NeuroImage, 12(1), 20-27.

Kouchaki, M., & Smith, I. H. (2014) The morning morality effect the influence of time of day on unethical behavior. Psychological Science, 25(1), 95-102.

Kozel, F., Johnson, K., Grenesko, E., Laken, S., Kose, S., Lu, X., Pollina, D., Ryan, A., & George, M. (2009a) Functional MRI detection of deception after committing a mock sabotage crime. Journal of Forensic Sciences, 54(1), 220-231.

Kozel, F. A., Laken, S. J., Johnson, K. A., Boren, B., Mapes, K. S., Morgan, P. S., & George, M. S. (2009b) Replication of functional MRI detection of deception. The Open Forensic Science Journal, 2, 6-11.

Kuss, K., Falk, A., Trautner, P., Elger, C. E., Weber, B., & Fliessbach, K. (2013) A reward prediction error for charitable donations reveals outcome orientation of donators. Social Cognitive and Affective Neuroscience, 8(2), 216-223.

Lamm, C., Decety, J., & Singer, T. (2011) Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. NeuroImage, 54(3), 2492-2502.

Lamm, C., & Singer, T. (2010) The role of anterior insular cortex in social emotions. Brain Structure and Function, 214(5), 579-591.

Lang, P. J. (1980) 'Behavioral treatment and bio-behavioral assessment: Computer applications', in J. J. J.Sidowski, T.Williams (ed), *Technology in Mental Health Care Delivery Systems*, New Jersey, Norwood.

Langleben, D., Loughead, J., Bilker, W., Ruparel, K., Childress, A., Busch, S., & Gur, R. (2005) Telling truth from lie in individual subjects with fast event-related fMRI. Human Brain Mapping, 26(4), 262-272.

Langleben, D., Schroeder, L., Maldjian, J., Gur, R., McDonald, S., Ragland, J., O'Brien, C., & Childress, A. (2002) Brain activity during simulated deception: an event-related functional magnetic resonance study. NeuroImage, 15(3), 727-732.

LeDoux, J. E. (2000) Emotion Circuits in the Brain. Annual Review of Neuroscience, 23(1), 155-184.

Lee, T., Au, R., Liu, H., Ting, K., Huang, C., & Chan, C. (2009) Are errors differentiable from deceptive responses when feigning memory impairment? An fMRI study. Brain and Cognition, 69(2), 406-412.

Lee, T., Lee, T., Raine, A., Chan, C., & Manzoni, O. (2010) Lying about the Valence of Affective Pictures: An fMRI Study. PloS one, 5(8), 244-245.

Lee, T., Liu, H., Chan, C., Ng, Y., Fox, P., & Gao, J. (2005) Neural correlates of feigned memory impairment. NeuroImage, 28(2), 305-313.

Lee, T., Liu, H., Tan, L., Chan, C., Mahankali, S., Feng, C., Hou, J., Fox, P., & Gao, J. (2002) Lie detection by functional magnetic resonance imaging. Human Brain Mapping, 15(3), 157-164.

Leotti, L. A., & Delgado, M. R. (2011) The inherent reward of choice. Psychological Science, 22(10), 1310-1318.

Leotti, L. A., Iyengar, S. S., & Ochsner, K. N. (2010) Born to choose: The origins and value of the need for control. Trends in Cognitive Sciences, 14(10), 457-463.

Levine, E. E., & Schweitzer, M. E. (2014) Are liars ethical? On the tension between benevolence and honesty. Journal of Experimental Social Psychology, 53, 107-117.

Levine, E. E., & Schweitzer, M. E. (2015) Prosocial lies: When deception breeds trust. Organizational Behavior and Human Decision Processes, 126, 88-106.

Lewis, A., Bardis, A., Flint, C., Mason, C., Smith, N., Tickle, C., & Zinser, J. (2012) Drawing the line somewhere: An experimental study of moral compromise. Journal of Economic Psychology, 33(4), 718-725.

Lieberman, M. D., & Cunningham, W. A. (2009) Type I and Type II error concerns in fMRI research: re-balancing the scale. Social Cognitive and Affective Neuroscience, 4(4), 423-428.

Liljeholm, M., Dunne, S., & O'Doherty, J. P. (2014) Anterior insula activity reflects the effects of intentionality on the anticipation of aversive stimulation. The Journal of Neuroscience, 34(34), 11339-11348.

Lindquist, M. A., & Wager, T. D. (2014) Principles of functional Magnetic Resonance Imaging. Handbook of Neuroimaging Data Analysis, London: Chapman & Hall.

Linke, J., Kirsch, P., King, A. V., Gass, A., Hennerici, M. G., Bongers, A., & Wessa, M. (2010) Motivational orientation modulates the neural response to reward. NeuroImage, 49(3), 2618-2625.

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.

López-Pérez, R., & Spiegelman, E. (2013) Why do people tell the truth? Experimental evidence for pure lie aversion. Experimental Economics, 16(3), 233-247.

Luan Phan, K., Magalhaes, A., Ziemlewicz, T., Fitzgerald, D., Green, C., & Smith, W. (2005) Neural correlates of telling lies: A functional magnetic resonance imaging study at 4 Tesla. Academic Radiology, 12(2), 164-172.

Lundquist, T., Ellingsen, T., Gribbe, E., & Johannesson, M. (2009) The aversion to lying. Journal of Economic Behavior & Organization, 70(1), 81-92.

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. NeuroImage, 19(3), 1233-1239.

Malti, T., & Krettenauer, T. (2013) The relation of moral emotion attributions to prosocial and antisocial behavior: A meta-analysis. Child Development, 84(2), 397-412.

Masip, J., Garrido, E., & Herrero, C. (2004) Defining deception. Anales de Psicología/Annals of Psychology, 20(1), 147-172.

Mason, W. A., Capitanio, J. P., Machado, C. J., Mendoza, S. P., & Amaral, D. G. (2006) Amygdalectomy and responsiveness to novelty in rhesus monkeys (Macaca mulatta): generality and individual consistency of effects. Emotion, 6(1), 73-81.

Masten, C. L., Morelli, S. A., & Eisenberger, N. I. (2011) An fMRI investigation of empathy for 'social pain'and subsequent prosocial behavior. NeuroImage, 55, 381-388.

Mazar, N., Amir, O., & Ariely, D. (2008) The dishonesty of honest people: A theory of self-concept maintenance. Journal of marketing research, 45(6), 633-644.

Mazar, N., & Ariely, D. (2006) Dishonesty in everyday life and its policy implications. Journal of Public Policy & Marketing, 25(1), 117-126.

McPherson, B., McMahon, K., Wilson, W., & Copland, D. (2011) "I know you can hear me": Neural correlates of feigned hearing loss. Human Brain Mapping, 33(8), 1964-1972.

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., & Ariely, D. (2009) Too tired to tell the truth: Self-control resource depletion and dishonesty. Journal of Experimental Social Psychology, 45(3), 594-597.

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016) Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. Psychophysiology, 53(5), 593-604.

Miller, G. R. (1983) Telling it like it isn't and not telling it like it is: Some thoughts on deceptive communication. The Jensen lectures: Contemporary communication studies, 91-116.

Mohamed, F. B., Faro, S. H., Gordon, N. J., Platek, S. M., Ahmad, H., & Williams, J. M. (2006) Brain Mapping of Deception and Truth Telling about an Ecologically Valid Situation: Functional MR Imaging and Polygraph Investigation—Initial Experience. Radiology, 238(2), 679-688.

Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002a) Functional networks in emotional moral and nonmoral social judgments. NeuroImage, 16(3), 696-703.

Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J. n., Andreiuolo, P. A., & Pessoa, L. (2002b) The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. The Journal of Neuroscience, 22(7), 2730-2736.

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006) Human fronto–mesolimbic networks guide decisions about charitable donation. Proceedings of the National Academy of Sciences, 103(42), 15623-15628.

Moll, J., Oliveira - Souza, D., & Zahn, R. (2008) The neural basis of moral cognition. Annals of the New York Academy of Sciences, 1124(1), 161-180.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005) The neural basis of human moral cognition. Nature Reviews Neuroscience, 6(10), 799-809.

Morrison, S. E., & Salzman, C. D. (2010) Re-valuing the amygdala. Current Opinion in Neurobiology, 20(2), 221-230.

Murray, E. A. (2007) The amygdala, reward and emotion. Trends in Cognitive Sciences, 11(11), 489-497.

Naqvi, N., Shiv, B., & Bechara, A. (2006) The role of emotion in decision making a cognitive neuroscience perspective. Current Directions in Psychological Science, 15(5), 260-264.

Nitschke, J. B., Sarinopoulos, I., Mackiewicz, K. L., Schaefer, H. S., & Davidson, R. J. (2006) Functional neuroanatomy of aversion and its anticipation. NeuroImage, 29(1), 106-116.

Nunez, J., Casey, B., Egner, T., Hare, T., & Hirsch, J. (2005) Intentional false responding shares neural substrates with response conflict and cognitive control. NeuroImage, 25(1), 267-277.

O'Doherty, J. (2004) Reward representations and reward-related learning in the human brain: insights from neuroimaging. Current Opinion in Neurobiology, 14(6), 769-776.

O'Doherty, J., Critchley, H., Deichmann, R., & Dolan, R. (2003) Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. Journal of Neuroscience, 23(21), 7931-7939.

Ogawa, S., Lee, T.-M., Nayak, A. S., & Glynn, P. (1990) Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. Magnetic Resonance in Medicine, 14(1), 68-78.

Öhman, A., & Mineka, S. (2001) Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. Psychological review, 108(3), 483-522.

Paulus, M. P., Rogalsky, C., Simmons, A., Feinstein, J. S., & Stein, M. B. (2003) Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. NeuroImage, 19(4), 1439-1448.

Pelphrey, K. A., Morris, J. P., & Mccarthy, G. (2004) Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. Journal of Cognitive Neuroscience, 16(10), 1706-1716.

Pessoa, L. (2009) How do emotion and motivation direct executive control? Trends in Cognitive Sciences, 13(4), 160-166.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002) Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. NeuroImage, 16(2), 331-348.

Phelps, E. A. (2009) Lying outside the laboratory: The impact of imagery and emotion on the neural circuitry of lie detection. Using Imaging to Identify Deceit: Scientific and Ethical Questions, 14-22.

Phelps, E. A., O'Connor, K. J., Gatenby, J. C., Gore, J. C., Grillon, C., & Davis, M. (2001) Activation of the left amygdala to a cognitive representation of fear. Nature Neuroscience, 4(4), 437-441.

Ploner, M., & Regner, T. (2013) Self-image and moral balancing: An experimental analysis. Journal of Economic Behavior & Organization, 93, 374-383.

Poldrack, R. A. (2006) Can cognitive processes be inferred from neuroimaging data? Trends in Cognitive Sciences, 10(2), 59-63.

Poldrack, R. A. (2007) Region of interest analysis for fMRI. Social Cognitive and Affective Neuroscience, 2(1), 67-70.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011) *Handbook of functional MRI data analysis*, Cambridge University Press.

Rilling, J. K., & Sanfey, A. G. (2011) The neuroscience of social decision-making. Annual Review of Psychology, 62, 23-48.

Robertson, D., Snarey, J., Ousley, O., Harenski, K., Bowman, F. D., Gilkey, R., & Kilts, C. (2007) The neural processing of moral sensitivity to issues of justice and care. Neuropsychologia, 45(4), 755-766.

Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014) Let's be honest: A review of experimental evidence of honesty and truth-telling. Journal of Economic Psychology, 45, 181-196.

Sabatinelli, D., Bradley, M. M., Lang, P. J., Costa, V. D., & Versace, F. (2007) Pleasure rather than salience activates human nucleus accumbens and medial prefrontal cortex. Journal of Neurophysiology (Bethesda), 98(3), 1374-1379.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003) The neural basis of economic decision-making in the ultimatum game. Science, 300(5626), 1755-1758.

Schauer, F. (2010) Neuroscience, Lie-Detection, and the law: Contrary to the prevailing view, the suitability of brain-based lie-detection for courtroom or forensic use should be determined according to legal and not scientific standards. Trends in Cognitive Sciences, 14(3), 101-103.

Sevinc, G., & Spreng, R. N. (2014) Contextual and perceptual brain processes underlying moral cognition: a quantitative meta-analysis of moral reasoning and moral emotions. PloS one, 9(2), e87427.

Shalvi, S., Dana, J., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011a) Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. Organizational Behavior and Human Decision Processes, 115(2), 181-190.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012) Honesty requires time (and lack of justifications). Psychological Science, 23(10), 1264-1270.

Shalvi, S., Handgraaf, M. J. J., & De Dreu, C. K. W. (2010) Ethical Manoeuvring: Why People Avoid Both Major and Minor Lies. British Journal of Management, 22(s1), S16-S27.

Shalvi, S., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011b) People avoid situations that enable them to deceive others. Journal of Experimental Social Psychology, 47(6), 1096-1106.

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004a) Brain responses to the acquired moral status of faces. Neuron, 41(4), 653-662.

Singer, T., & Lamm, C. (2009) The social neuroscience of empathy. Annals of the New York Academy of Sciences, 1156(1), 81-96.

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004b) Empathy for pain involves the affective but not sensory components of pain. Science, 303(5661), 1157-1162.

Sip, K., Lynge, M., Wallentin, M., McGregor, W., Frith, C., & Roepstorff, A. (2010) The production and detection of deception in an interactive game. Neuropsychologia, 48(12), 3619-3626.

Sip, K., Roepstorff, A., McGregor, W., & Frith, C. (2008) Detecting deception: the scope and limits. Trends in Cognitive Sciences, 12(2), 48-53.

Sip, K. E., Skewes, J. C., Marchant, J. L., McGregor, W. B., Roepstorff, A., & Frith, C. D. (2012) What if I get busted? Deception, choice, and decision-making in social interaction. Frontiers in Neuroscience, 6(58), 1-10.

Spence, S. (2004) The deceptive brain. Journal of the Royal Society of Medicine, 97(1), 6-9.

Spence, S., Farrow, T., Herford, A., Wilkinson, I., Zheng, Y., & Woodruff, P. (2001) Behavioural and functional anatomical correlates of deception in humans. Neuroreport, 12(13), 2849-2853.

Spence, S., Hunter, M., Farrow, T., Green, R., Leung, D., Hughes, C., & Ganesan, V. (2004) A cognitive neurobiological account of deception: evidence from functional neuroimaging. Philosophical Transactions of the Royal Society B: Biological Sciences, 359(1451), 1755-1762.

Spence, S., Kaylor-Hughes, C., Farrow, T., & Wilkinson, I. (2008) Speaking of secrets and lies: the contribution of ventrolateral prefrontal cortex to vocal deception. NeuroImage, 40(3), 1411-1418.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., & Fehr, E. (2007) The neural signature of social norm compliance. Neuron, 56(1), 185-196.

Spreckelmeyer, K. N., Krach, S., Kohls, G., Rademacher, L., Irmak, A., Konrad, K., Kircher, T., & Gründer, G. (2009) Anticipation of monetary and social reward differently activates mesolimbic brain structures in men and women. Social Cognitive and Affective Neuroscience, 4(2), 158-165.

Sprengelmeyer, R. (2007) The neurology of disgust. Brain, 130(7), 1715-1717.

Sprengelmeyer, R., Rausch, M., Eysel, U. T., & Przuntek, H. (1998) Neural structures associated with recognition of facial expressions of basic emotions. Proceedings of the Royal Society of London B: Biological Sciences, 265(1409), 1927-1931.

Sun, D., Chan, C. C., Hu, Y., Wang, Z., & Lee, T. M. (2015a) Neural Correlates of outcome processing post dishonest choice: An fMRI and ERP study. Neuropsychologia, 68, 148-157.

Sun, D., Lee, T. M., & Chan, C. C. (2015b) Unfolding the spatial and temporal neural processing of lying about face familiarity. Cerebral Cortex, 25(4), 927-936.

Sutter, M. (2009) Deception through telling the truth?! experimental evidence from individuals and teams. The Economic Journal, 119(534), 47-60.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008) The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychological Science, 19(4), 339-347.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007) Moral emotions and moral behavior. Annual Review of Psychology, 58, 345-372.

Tangney, J. P., Wagner, P., & Gramzow, R. (1992) Proneness to shame, proneness to guilt, and psychopathology. Journal of Abnormal Psychology, 101(3), 469-478.

Thielscher, A., & Pessoa, L. (2007) Neural correlates of perceptual choice and decision making during fear–disgust discrimination. The Journal of Neuroscience, 27(11), 2908-2917.

Valdesolo, P., & DeSteno, D. (2006) Manipulations of emotional context shape moral judgment. Psychological Science, 17(6), 476-477.

Verschuere, B., Spruyt, A., Meijer, E. H., & Otgaar, H. (2011) The ease of lying. Consciousness and cognition, 20(3), 908-911.

Vincent, J. L., Kahn, I., Snyder, A. Z., Raichle, M. E., & Buckner, R. L. (2008) Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. Journal Of Neurophysiology, 100(6), 3328-3342.

Volz, K. G., Vogeley, K., Tittgemeyer, M., von Cramon, D. Y., & Sutter, M. (2015) The neural basis of deception in strategic interactions. Frontiers in behavioral neuroscience, 9(27), 1-12.

Vrij, A. (2007) Deception: A social lubricant and a selfish act. Frontiers of social psychology: Social communication, 309-342.

Vrij, A. (2008) *Detecting lies and deceit: Pitfalls and opportunities*, Wiley-Interscience.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006) Detecting deception by manipulating cognitive load. Trends in Cognitive Sciences, 10(4), 141-142.

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008) Increasing cognitive load to facilitate lie detection: the benefit of recalling an event in reverse order. Law and Human Behavior, 32(3), 253-265.

Walczyk, J. J., Mahoney, K. T., Doverspike, D., & Griffith-Ross, D. A. (2009) Cognitive lie detection: Response time and consistency of answers as cues to deception. Journal of Business and Psychology, 24(1), 33-49.

Walczyk, J. J., Roper, K. S., Seemann, E., & Humphrey, A. M. (2003) Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. Applied Cognitive Psychology, 17(7), 755-774.

Wiltermuth, S. S. (2011) Cheating more when the spoils are split. Organizational Behavior and Human Decision Processes, 115(2), 157-168.

Wright, G. R., Berry, C. J., & Bird, G. (2013) Deceptively simple… The "deception-general" ability and the need to put the liar under the spotlight. Frontiers in Neuroscience, 7(152), 1-9.

Wu, D., Loke, I. C., Xu, F., & Lee, K. (2011) Neural correlates of evaluations of lying and truth-telling in different social contexts. Brain Research, 1389, 115-124.

Xu, F., Evans, A. D., Li, C., Li, Q., Heyman, G., & Lee, K. (2013) The role of honesty and benevolence in children's judgments of trustworthiness. International Journal of Behavioral Development, 37(3), 257-265.

Young, K. D., Zotev, V., Phillips, R., Misaki, M., Yuan, H., Drevets, W. C., & Bodurka, J. (2014) Real-time FMRI neurofeedback training of amygdala activity in patients with major depressive disorder. 9(2), e88785.

Yu, R., Calder, A. J., & Mobbs, D. (2014) Overlapping and distinct representations of advantageous and disadvantageous inequality. Human Brain Mapping, 35(7), 3290-3301.

Zhu, L., Jenkins, A. C., Set, E., Scabini, D., Knight, R. T., Chiu, P. H., King-Casas, B., & Hsu, M. (2014) Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. Nature neuroscience, 17(10), 1319-1321.

List of Figures

List of Tables