Lies that feel honest: Dissociating between incentive and deviance processing when evaluating dishonesty

Lelieveld, G.-J.; Shalvi, S.; Crone, E.A.

DOI
10.1016/j.biopsycho.2016.03.009

Publication date
2016

Document Version
Final published version

Published in
Biological Psychology

License
Article 25fa Dutch Copyright Act

Link to publication

Citation for published version (APA):
Lelieveld, G-J., Shalvi, S., & Crone, E. A. (2016). Lies that feel honest: Dissociating between incentive and deviance processing when evaluating dishonesty. *Biological Psychology*, *117*, 100-107. https://doi.org/10.1016/j.biopsycho.2016.03.009

Biological Psychology 117 (2016) 100–107

Contents lists available at ScienceDirect

Biological Psychology

journal homepage: www.elsevier.com/locate/biopsycho

Lies that feel honest: Dissociating between incentive and deviance processing when evaluating dishonesty

Gert-Jan Lelieveld[a,b,*], Shaul Shalvi[c], Eveline A. Crone[a,b,d]

[a] Institute of Psychology, Leiden University, The Netherlands
[b] Leiden Institute for Brain and Cognition, Leiden, The Netherlands
[c] CREED, Faculty of Economics and Business, University of Amsterdam, The Netherlands
[d] Department of Psychology, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 4 September 2015
Received in revised form 8 March 2016
Accepted 16 March 2016
Available online 17 March 2016

Keywords:
Lies
Evaluation
Justifications
fMRI
Dorsal ACC
Conflict

ABSTRACT

This study investigated neural responses to evaluations of lies made by others. Participants learned about other individuals who were instructed to privately roll a die twice and report the outcome of the first roll to determine their pay (with higher rolls leading to higher pay). Participants evaluated three types of outcomes: honest reports, justifiable lies (reporting the second outcome instead of the first), or unjustifiable lies (reporting a different outcome than both die rolls). Evaluating lies relative to honest reports was associated with increased activation in the anterior cingulate cortex (ACC), insula and lateral prefrontal cortex. Moreover, justifiable lies were associated with even stronger activity in the dorsal ACC and dorsolateral prefrontal cortex compared to unjustifiable lies. These activities were more pronounced for justifiable lies where the deviance from the real outcome was larger. Together, these findings have implications for understanding how humans judge misconduct behavior of others.

1. Introduction

The way individuals judge dishonesty is a central aspect of daily interactions. It allows juries to craft verdicts, inform partners to stay together or get divorced, and tax agents to initiate financial investigations of suspected firms (Levine, 2010). When judging dishonest behavior, certain contextual factors can cause people to consider certain lies as less dishonest than other lies (e.g., when people lie to help another person, see Gino, Ayal, & Ariely, 2013). This situation, in which it is not right away evident if the behavior should be judged as clearly dishonest, is thought to result in cognitive conflict. We define cognitive conflict as those situations that require a selection among a set of equally permissible responses (see Botvinick, 2007). In the current study, we hypothesized that individuals experience more conflict when evaluating lies that can be justified, compared to evaluating lies that cannot be justified, given that there will be more competition between right/wrong selection processes. We tested this hypothesis using functional neuroimaging, which allowed us to test conflict experience at the neural level (Botvinick, 2007).

Prior research on dishonest behavior mainly focused on the behavior of the lie-teller. This research has shown that when motivated to do so, individuals often lie for financial profit (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996), but at the same time, people want to maintain an honest self-concept (Mazar, Amir, & Ariely, 2008). Thus, there is a balance between justifiable and non-justifiable lie telling, although this balance may differ between individuals and situations. It is a well-replicated finding that individuals restrict their dishonesty and lie more often for small amounts than for large amounts (Ayal & Gino, 2011; Fischbacher & Follmi-Heusi, 2008). Less is known, however, about whether a similar distinction is made when individuals evaluate lies of others. Although there is a large body of research that focuses on lie detection and the accuracy thereof (for a review see Rosenfeld, Ben-Shakhar, & Ganis, 2016), hardly any research has focused on how individuals evaluate whether and under what circumstances lies of others are justifiable or not. This is an important issue given that even small or justifiable lies may accumulate to large societal costs. After all, on aggregate many little lies pile up to a hefty sum (Ariely, 2012).

One method that allows us to gain a deeper understanding of how individuals evaluate lies of others is by using neuroimaging. Prior research has shown that in general moral judgment makes use of brain regions dedicated to social cognition (Greene & Haidt, 2002). These brain regions include, for example, the orbitofrontal

* Corresponding author at: Institute of Psychology, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands.
E-mail address: lelieveldgj@fsw.leidenuniv.nl (G.-J. Lelieveld).

http://dx.doi.org/10.1016/j.biopsycho.2016.03.009
0301-0511/© 2016 Elsevier B.V. All rights reserved.

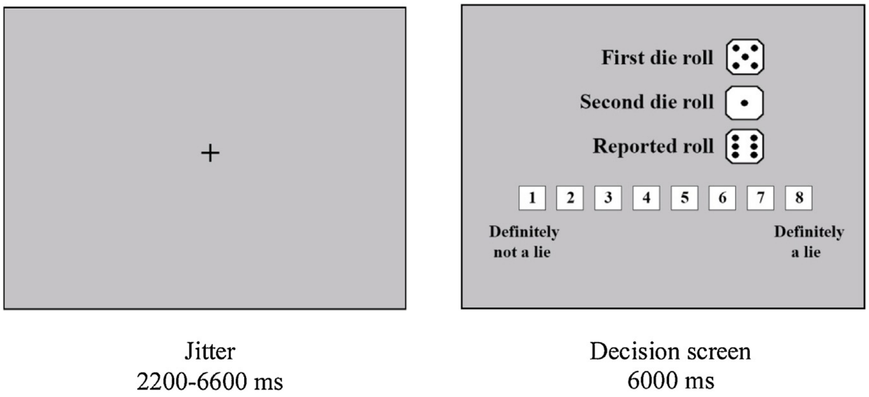

Fig. 1. Visual display and timing of the events in the scanner task in milliseconds (ms). After a jittered fixation cross, a screen displayed the two die rolls and the reported roll (here an example of the first two rolls are "5" and "1", and the reported roll is "6"). We measured activation at the onset of this decision screen. Participants had a maximum response time of 6000 ms. After the response, the decision screen remained on the screen until 6000 ms after the onset of the decision screen.

(OFC) and medial prefrontal cortex (mPFC) (Moll et al., 2002). However, besides these brain regions, the evaluation of lies is also thought to be associated with cognitive processes, as is evident from prior studies which reported increased activity in the dorsolateral prefrontal cortex (dlPFC) and the dorsal anterior cingulate cortex (ACC) when evaluating moral dilemmas (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004). In the current study our first aim was to test whether the same regions that have previously been associated with judging moral conflict (the OFC and mPFC) and cognitive control (the dlPFC and ACC) are also involved in the evaluation of lies of others, by comparing evaluations of lies to evaluations of honest reports, using a die roll paradigm.

A second aim of the current study was to investigate the moderating role of justifications on the evaluations of lies. An important determinant for how lies of others are evaluated is the extent to which lies can be justified (Schweitzer & Hsee, 2002; Shalvi, Gino, Barkan, & Ayal, 2015). People clearly restrict their dishonesty and seem to stretch the truth to the extent they can justify such behavior (Toure-Tillery & Fishbach, 2012). This pattern is also well documented in the justifiable lies paradigm (Shalvi, Dana, Handgraaf, & De Dreu, 2011). In this paradigm, participants are asked to roll a die privately and earn money as a function of their reported die roll outcome (1 = $1, 2 = $2, etc.). Since only participants see their die rolls, they can inflate their reports and leave the experiment richer than when they had reported honestly. In the experimental condition participants were asked to roll the die three times, check the outcome of each roll, but then report the outcome of the first roll only (see Fischbacher & Follmi-Heusi, 2008). In the control condition, holding all other aspects constant, participants were asked to roll only once before reporting their outcome. Results showed that participants who rolled three times lied more often than those rolling only once, because participants justified their lie by reporting the better outcome on the second or third roll (see also Gino & Ariely, 2012; Shalvi, Eldar, & Bereby-Meyer, 2012).

In a second series of experiments, Shalvi and Leiser (2013) investigated how other individuals judged these lies, thereby testing the evaluation of dishonesty. Participants were presented with a scenario describing the behavior of other participants involved in the die rolling experiment described above. Participants were asked to rank the extent to which they found each of the presented combinations to be a lie (1 = *not at all* to 6 = *very much*). They presented die roll combinations (i.e., the outcome of the first and second roll and the reported outcome) that were either honest (1st roll = report), or dishonest (1st roll < report). Critically, within the dishonest combinations, some lies could be justified by reporting the outcome of the (irrelevant for pay) second roll (justifiable combinations). Other participants were presented with lies that could not be justified, where a higher outcome was reported that did not match any of the die rolls (unjustifiable combinations). Participants judged the justifiable combinations as less of a lie compared to the unjustifiable combinations. This paradigm provides a valuable context for examining how individuals evaluate lies, and more specifically, if lies that can be justified are experienced as less dishonest. Possibly, the evaluation of justifiable lies elicits more conflict than the evaluation of unjustifiable lies, because participants may have difficulty deciding whether this is a complete lie or not.

To examine these conflict responses, we made use of neuroimaging to test cognitive conflict processes in more detail. Based on prior studies, it is well documented that the experience of cognitive conflict is associated with activity in the dorsal ACC and the dlPFC (Greene et al., 2004; Hayashi et al., 2014). These regions are well known for their role in signaling conflict and adjusting behavior to changing environmental cues (Botvinick, 2007; Botvinick, Cohen, & Carter, 2004; Van Veen & Carter, 2006). Conflict-related activity in the ACC is not restricted to behavioral conflict, such as when response mappings are competing (Carter & Van Veen, 2007), but is also found when experiencing social conflict, such as cognitive dissonance (Van Veen, Krug, Schooler, & Carter, 2009) or social expectation violations (Somerville, Heatherton, & Kelley, 2006). In the current study, we tested if justifiable lies elicited more activity in the ACC and dlPFC, under the hypothesis that justifiable lies create more conflict.

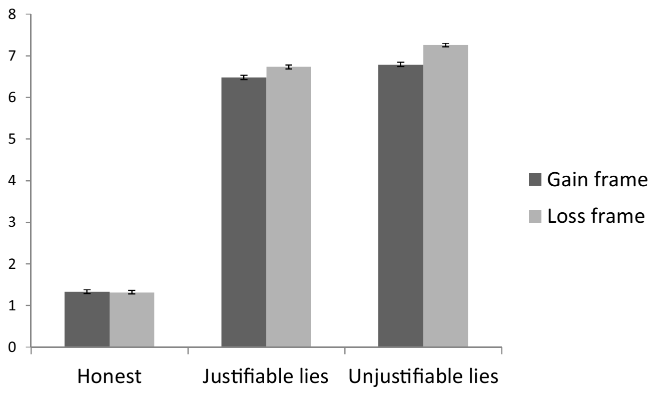

Fig. 2. Behavioral results for participants' evaluations, as a function of die report and frame. Larger means represent harsher evaluations (i.e., participants evaluated these report as more of a lie). Error bars represent standard errors calculated according to the method of Loftus and Masson (1994), see also Pfister & Janczyk, 2013).

Table 1
Brain regions revealed by whole brain contrasts, including MNI coordinates. Peak voxels reported at $p < .001$ uncorrected, at least 10 contiguous voxels (voxels size was $3.0 \times 3.0 \times 3.0$ mm).

Anatomical region	L/R	Voxels	Z	MNI coordinates		
				x	y	z
3 × 2 ANOVA—Main effect of die reports						
Dorsal Anterior Cingulate Cortex	L/R	38	3.73	−6	23	40
			3.23	9	20	37
Premotor Cortex	L	19	3.84	−24	−1	55
3 × 2 ANOVA—Main effect of frame						
Dorsal Anterior Cingulate Cortex	L/R	246	5.02	9	14	37*
			4.33	−27	−1	46*
			4.28	−30	5	49*
Bilateral Prefrontal Cortex	L	63	4.37	−33	50	1*
			3.71	−27	38	10*
			3.55	−27	56	7*
	R	120	4.92	21	56	−2*
			3.89	33	47	4*
			3.72	9	59	−5*
Justifiable Lies > Unjustifiable Lies						
Dorsal Anterior Cingulate Cortex	L/R	53	3.81	3	26	19
			3.77	−3	23	22
			3.57	3	32	13
Medial Prefrontal Cortex	R	16	3.86	9	41	−5
Striatum	R	16	3.80	21	17	−2
	L	12	3.75	−24	23	7
Justifiable Lies > Honest Reports						
Dorsal Anterior Cingulate Cortex	L/R	6401	8.63	0	17	46*
Striatum			5.50	−15	8	−2*
Prefrontal Cortex			5.55	−24	26	4*
Inferior Parietal Lobule	L/R	2213	6.89	12	−70	52*
			5.75	45	−34	43*
			5.57	42	−37	40*
Unjustifiable Lies > Honest Reports						
Dorsal Anterior Cingulate Cortex	L/R	1450	5.79	−3	14	49*
Striatum		115	3.85	12	20	1*
Prefrontal Cortex		830	5.02	−39	5	28*
Inferior Parietal Lobule	L/R	1501	5.96	−15	−70	55*
			5.71	−21	−70	49*
			5.46	45	−37	40*
(Justifiable Lies > Unjustifiable Lies) large deviance > (Justifiable Lies > Unjustifiable Lies) small deviance						
Dorsolateral Prefrontal Cortex	R	14	3.61	30	53	28
			3.52	27	56	25
Justifiable Lies > Unjustifiable Lies for large deviance						
Dorsal Anterior Cingulate Cortex	L/R	69	4.07	−3	26	25
			3.45	6	32	22
			3.35	−9	32	22
Dorsolateral Prefrontal Cortex	R	38	3.87	27	44	34
			3.79	30	50	28

* The results remained significant with an FDR-corrected threshold of $p < .05$, with an extent threshold of 10 continuous voxels.

Crucially, we investigated which processes might drive this increase in activity. For this purpose, we used fMRI to examine two hypotheses with respect to the circumstances under which individuals experience the evaluation of justifiable lies as more or less conflicting. The first hypothesis states that individuals evaluate justifiable lies of others dependent on the incentives that are associated with lying. Lies are less acceptable when it is meant for personal gain than in the face of a potential loss (Kern & Chugh, 2009; see also Tversky & Kahneman, 1981). Possibly individuals experience more conflict when they evaluate justifiable lies meant for personal gain than when they evaluate justifiable lies when the other person tries to avoid losses (Kern & Chugh, 2009; see also Tversky & Kahneman, 1981). There may thus be more competition between right/wrong selection processes for justified lies meant for personal gain. Therefore, we tested if framing the justifiable lie as a way to secure a monetary gain should result in stronger activity in the dorsal ACC and the dlPFC than framing the lie as a way to avoid monetary loss. This is referred to as the incentive hypothesis, because it involves an evaluation of incentives for others. The second hypothesis states that evaluating justifiable lies creates more conflict when they result in a large deviance from the truth compared to a small deviance. Under this hypothesis, when these outcomes are more deviant, this should result in increased activity in the dorsal ACC and dlPFC independent of whether the deviation results in loss or gain. This hypothesis is referred to as the deviance hypothesis.

To this end, we asked participants to complete the justifiable lies task while fMRI images were collected, in the context of evaluating lies of others. We manipulated whether lies from others (justifiable or not) were used as a mean to secure a gain vs. to avoid a loss. Secondly, we manipulated the deviance between outcomes of the first and second die roll (e.g., reporting 2 instead of 1 versus reporting 6 instead of 1). This design allowed us to test the role of incentives and deviance in the same experimental design, and thereby will provide a rigorous test of the brain regions associated with evaluating justifiable lies.

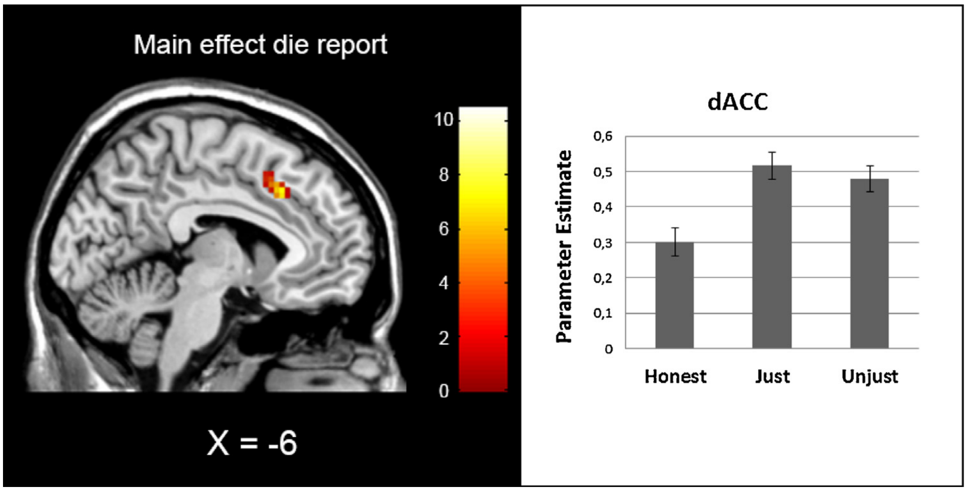

Fig. 3. Brain regions of main effect of report in the 3 (report) × 2 (frame) full factorial design (threshold at $p < .001$, uncorrected). Activation was detected in the dACC (MNI coordinates: x = −6, y = 23, z = 40). Error bars represent standard errors calculated according to the method of Loftus and Masson (1994).

2. Method

2.1. Participants

Based on previous neuroimaging studies investigating the evaluation of lies (Hayashi et al., 2014; Wu, Loke, Xu, & Lee, 2011), we aimed to include a minimum of 20 participants per between-subjects condition. The final sample consisted of forty-five healthy right-handed paid volunteers, who were students from Leiden University. Due to a technical error during the scan session the data from one participant were lost. We therefore analyzed the data from forty-four participants (25 female, 19 male; $M_{age} = 21.00$, $SD = 1.87$; age range 18–25). None had any history of neurological or psychiatric disorder based on self-report and all were medication-free. All participants gave written informed consent for the study, and all procedures were approved by the medical ethical committee of the Leiden University Medical Center (LUMC).

2.2. Experimental task

In the scanner participants evaluated others' honest reports vs. justifiable lies vs. unjustifiable lies. Prior to the scan session, participants received written instruction explaining the die roll task. Participants read about other students who were instructed to pri-

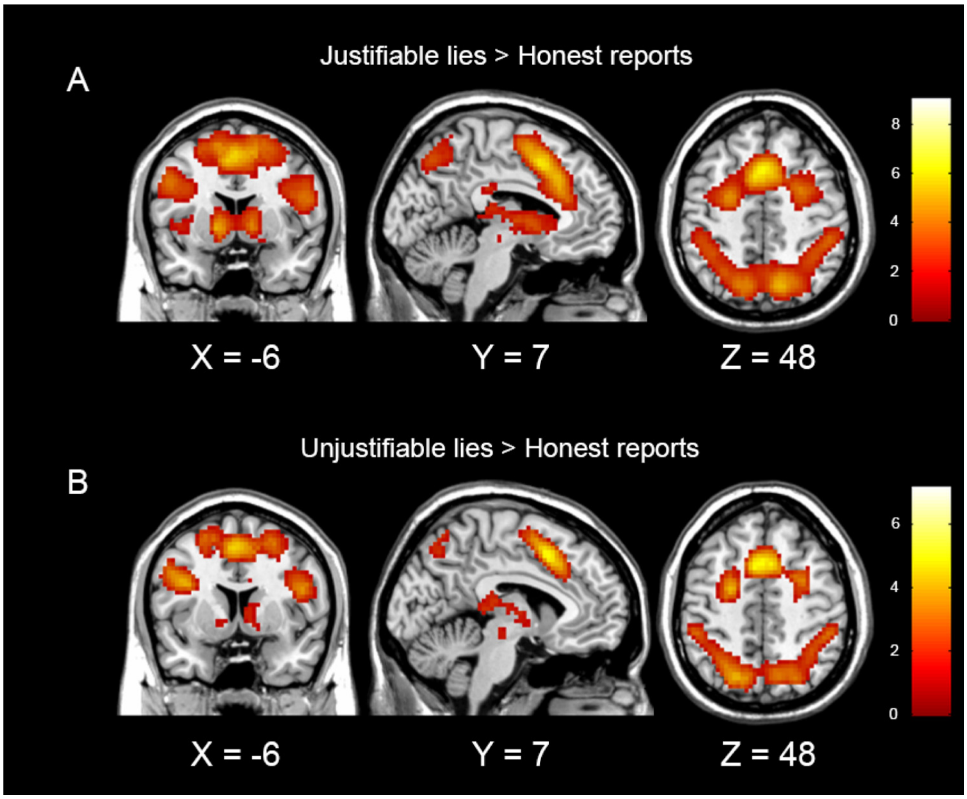
A
Justifiable lies > Honest reports
X = -6
Y = 7
Z = 48
8
6
4
2
0
B
Unjustifiable lies > Honest reports
X = -6
Y = 7
Z = 48
6
4
2
0

Fig. 4. Whole brain results for regions which were active in the (A) Justifiable Lies > Honest Reports contrast and the B) Unjustifiable Lies > Honest Reports. In both contrasts activation was detected in the ACC, anterior insula, dlPFC and several regions in the parietal lobule (thresholds at $p < .05$, FDR corrected).

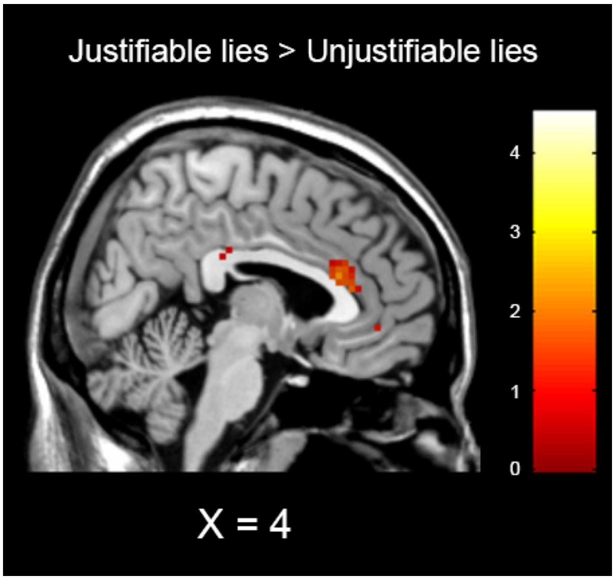

Fig. 5. Whole brain results for regions which were active in the Justifiable Lies > Unjustifiable Lies contrast (threshold at $p < .001$, uncorrected). Activation was detected in the ACC (MNI coordinates: x = 3, y = 26, z = 19).

vately roll a six-faced die twice and asked to report the outcome of the first roll to get paid (Shalvi et al., 2011). Since the students privately rolled the die and reported the outcome, they had an opportunity to lie. We distinguished between two types of lies. First, instead of reporting the outcome of the first roll, they could report the outcome of the second roll in case this outcome was higher (i.e., a justifiable lie). Second, they could report a higher number which they did not observe in the first or second roll (i.e., an unjustifiable lie).

In the scanner, participants were asked to evaluate the extent to which different combinations of rolls and reports were considered to be lies. Specifically, participants engaged in 60 trials: 20 trials in which the other person reported honestly (i.e., report = 1st roll outcome), 20 trials in which the other person used a justifiable lie (i.e., report = 2nd roll outcome > 1st roll outcome), and 20 trials in which the other person used an unjustifiable lie (i.e., report > both 1st and 2nd rolls). Each trial consisted of a jittered fixation cross (2200–6600 ms), followed by an evaluation screen containing the outcomes of the 1st and 2nd rolls, and the reported outcome (see Fig. 1). The evaluation screen remained on the screen for 6000 ms. During presentation of the evaluation screen, participants were asked to what extent they considered the report to be a lie on an 8-point-scale. The order of the 8-point scale (1 = *not at all*, and 8 = *very much*, versus 1 = *very much*, and 8 = *not at all*) was counterbalanced across participants. Responses were made with the left and right index finger, middle finger, ring finger, and little finger (i.e., they used 8 fingers for the 8-point scale) using response boxes attached to the upper legs. Participants' responses were made bold on the 8-point scale and remained on the screen for the entire 6000 ms. When participants did not respond in time, participants were notified (this lasted 1000 ms), after which the next trial started.

2.3. Task manipulations for incentives versus deviance

We manipulated how we framed the lie as well as the deviance of the lie. First, we manipulated (between subjects) whether participants' lies (justifiable or not) were used as a mean to secure a gain vs. to avoid a loss. Specifically, participants in the gain condition ($N = 22$; 13 females) learned that the students engaging in the die task earned money based on their reports: reporting 1 = €1, 2 = €2, etc. up to 6 = €6. Participants in the loss condition ($N = 22$; 12 females) learned that the students received an initial endowment of €7 and that they could lose money based on their reports. Reporting 1 meant losing €6 and thus ending up with €7 – €6 = €1, reporting 2 meant losing €5 and thus ending up with €7 – €5 = €2, etc. Thus, in both conditions a higher reported outcome meant higher payoff. However, in the positive framing condition the higher outcome secured a gain, whereas in the negative framing condition the higher outcome prevented a loss.

Second, within the justifiable and unjustifiable lies trials, we also manipulated the deviance of the lie to be small or large. Specifically, half of the lies in each block were minor (i.e., reporting an outcome which was larger than the 1st roll by 1 or 2) whereas the other half of the lies were larger (i.e., reporting an outcome which was larger than the 1st outcome by 3, 4, or 5).

Note that the above design is not fully factorial, since it is not possible to manipulate the deviance (of dishonesty) of honest reports. For this reason we used a 3 (die roll reports: honest vs. justifiable lies vs. unjustifiable lies) × 2 (frame: gain frame vs. loss frame) design to compare both types of lies to honest reports and to test the incentive hypothesis, and a 2 (die roll reports: justifiable lies vs. unjustifiable lies) × 2 (deviance: small vs. large

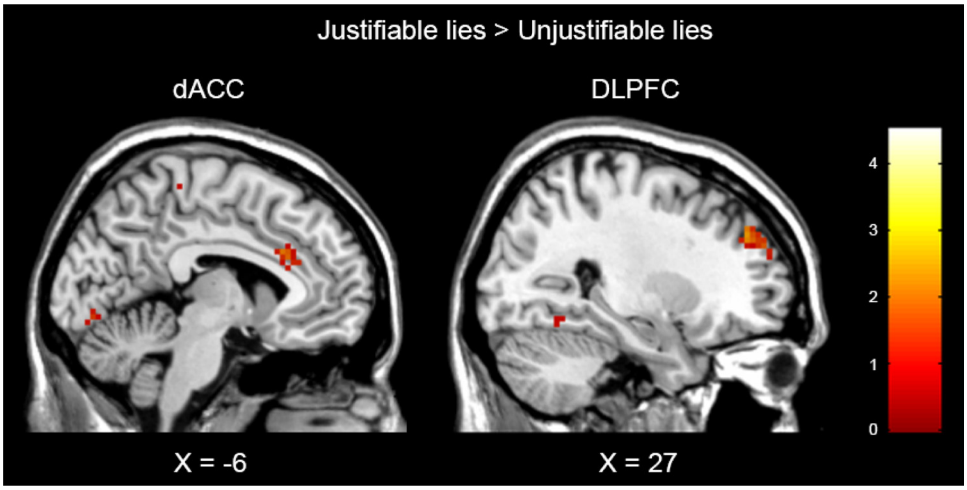

Fig. 6. Whole brain results for regions which were active in the Justifiable Lies > Unjustifiable Lies contrast for the large deviance lies only (threshold at $p < .001$, uncorrected). Activation was detected in the ACC (MNI coordinates: x = –3, y = 26, z = 25) and dlPFC (MNI coordinates: x = 30, y = 50, z = 28).

deviance) × 2 (frame: gain frame vs. loss frame) design to test the deviance hypothesis.

2.4. Procedure

Before being scanned, participants in both gain and loss conditions, completed a comprehension session assuring that they understood how much money the other person would earn when reporting each of the possible outcomes (1 to 6). These sessions assured all participants were able to quickly answer how much the person reporting would earn based on his/her reports. This procedure was carried out to minimize potential differences between the gain and loss conditions. After successfully completing the comprehension task, participants were introduced to the evaluation task and completed 12 trials on a computer situated outside the scanner room. They then were guided into the MRI scanner and performed the complete task. At the end of the task, participants were debriefed and received 20 euros for participation.

2.5. fMRI Data acquisition

Scanning was performed on a 3.0T Philips Achieva scanner at the LUMC. Functional data were acquired using a T2*-weighted echo-planar imaging (EPI) sequence (echo time/TE = 30 ms, repetition time/TR = 2200 ms, slice-matrix = 80 × 80, slice-thickness = 2.75 mm, slice gap = 0.28 mm gap, field of view [FOV] = 220 mm), during one fMRI run which lasted for approximately ten minutes. At the end of the scan session, a high-resolution T2-weighted high-resolution anatomical scan (same slice prescription as EPI) was collected.

2.6. fMRI Data analysis

Data pre-processing and analyses were conducted with SPM8 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm8) implemented in MATLAB (Mathworks, Sherborn, MA). All functional images were realigned and slice-time corrected using the middle slice as reference. Then they were spatially normalized to T1 templates and spatially smoothed with a Gaussian kernel (8 mm, full-width at half-maximum). A canonical haemodynamic response function (HRF) was convolved at the onset of the presentation of the die reports (impulse function: zero duration).

As noted above, we tested a 3 (evaluated die reports) × 2 (frame) design, and a 2 (evaluated die reports) × 2 (frame) × 2 (deviance) design with die reports and deviance as within and frame as between subjects factors. Analyses were carried out using the general linear model in SPM8. Contrast parameter images were computed for each individual. The resulting contrast images were submitted to second-level group analyses. At the group level, we performed an Analysis of Variance (ANOVA) as well as one-tailed *t*-tests on these images, treating participants as a random effect. Results were considered significant at an uncorrected threshold $p < .001$ with an extent threshold of ten continuous voxels (thresholds were based on recommendations from Lieberman & Cunningham (2009), to produce a desirable balance between Type I and Type II errors. Table 1 reports which results remained significant with an FDR $p < .05$, >10 contiguous voxels threshold).

We further extracted parameter estimates from the regions that were identified in the whole brain analyses using the MARSBAR toolbox for SPM8 (Brett, Anton, Valabreque, & Poline, 2002), to further characterize patterns of activity.

3. Results

3.1. Behavioral results

Participants' evaluation of the die reports were submitted to a 3 × 2 mixed-model Analysis of Variance (ANOVA) with die roll reports (honest vs. justifiable lies vs. unjustifiable lies) as repeated-measures variable and frame (gain frame vs. loss frame) as between-subjects variable. The analysis yielded only a main effect of die reports, $F(2, 84) = 530.62$, $p < .001$, $\eta^2 = .93$ (see Fig. 2). LSD post hoc tests showed that participants evaluated justifiable lies ($M = 6.61$, $SD = 1.06$) and unjustifiable lies ($M = 7.02$, $SD = 1.00$) as more of a lie than honest reports ($M = 1.33$, $SD = .78$; both $ps < .001$). Importantly, replicating previous findings (Shalvi et al., 2011; Shalvi & Leiser, 2013), the LSD post hoc tests further showed that justifiable lies were evaluated as less of a lie compared to unjustifiable lies ($p < .001$). The main effect of frame as well as the reports × frame interaction were not significant (p's > .20).

Next, we tested the effect of deviance of the lie. Focusing on evaluations of lies, a 2 × 2 × 2 mixed-model ANOVA with die report (justifiable lies vs. unjustifiable lies) and deviance (large vs. small) as repeated-measures within subjects variables and frame (gain frame vs. loss frame) as between-subjects variable, again revealed a main effect of die report, $F(1, 42) = 19.52$, $p < .001$, $\eta^2 = .32$, as well as a main effect of deviance, $F(1, 42) = 13.22$, $p < .001$, $\eta^2 = .24$. Participants evaluated large deviance lies as more of a lie ($M = 7.10$, $SD = .80$) than small deviance lies ($M = 6.56$, $SD = 1.31$). There was no three way interaction $F(1, 42) = .01$, $p = .94$, $\eta^2 = .00$, no interaction between die report and deviance $F(1, 42) = .04$, $p = .84$, $\eta^2 = .00$, and no interaction effects of frame and die report $F(1, 42) = 1.20$, $p = .28$, $\eta^2 = .03$ and of frame and deviance $F(1, 42) = 3.58$, $p = .07$, $\eta^2 = .08$.

3.2. fMRI Results

3.2.1. Evaluating (dis) honesty

To assess the differences in the BOLD signal depending on reports (honest vs. justifiable lies vs. unjustifiable lies) and frame (gain vs. loss) a 3 × 2 ANOVA was conducted at the whole brain level. The main effect of die reports resulted in activation in a cluster in the dACC, $F(2, 123) = 9.98$, $p < .001$ (see Fig. 3), and a cluster in the lateral PFC (see Table 1).

The analysis also resulted in a main effect of frame, which showed activation in several regions in the anterior cingulate and prefrontal cortex (see Table 1). However, the reports × frame interaction did not result in any significant regions, suggesting that the dACC activation for justifiable lies was present across frame contexts.

In order to further test the differences between the evaluations of the different die reports we performed separate one-tailed *t*-tests, averaged across the frame conditions. The contrasts Justifiable Lies > Honest Reports and Unjustifiable Lies > Honest Reports both showed activation in a wide network including the dACC, anterior insula and prefrontal cortex (see Fig. 4A and B, and Table 1). As expected, our critical test for comparing justifiable lies to unjustifiable lies resulted again in activation in the dACC (see Fig. 5 and Table 1).

3.2.2. Testing the deviance hypothesis

The 2 (die reports) × 2 (deviance) × 2 (frame) ANOVA did not show any significant main effects or interaction effects. However, this large ANOVA may have been underpowered to detect die report × deviance interactions. Based on the behavioral results, we compared the effects of justifiable and unjustifiable lies for the large and small deviance lies separately, as an explorative post hoc analysis. A two sample *t*-test comparing activation for the Justifiable Lies > Unjustifiable Lies contrast for High Deviance > Low Deviance showed increased activation in the dlPFC (see Table 1). Subsequently we compared the effects of justifiable and unjustifiable lies for large and small deviance lies separately. The Justifiable Lies > Unjustifiable Lies contrast for the large deviance lies showed significant activation in the dACC and dlPFC (see Fig. 6 and Table 1). The same contrast for the small deviance lies did not show any significant results. The reverse contrast also did not show significant activation for the large and small deviance lies.

4. Discussion

This study investigated which brain regions are associated with the evaluation of lies of others compared to the evaluation of honest reports, and secondly tested the experience of conflict when evaluating justifiable vs. unjustifiable lies of others. Behaviorally, the results replicate prior findings showing that lies with a similar deviance are evaluated differently, depending on whether they could be justified or not (Shalvi et al., 2011; Shalvi & Leiser, 2013). The neuroimaging results showed that when comparing the evaluation of lies to the evaluation of honest reports, there was increased activation in the dorsal ACC, anterior insula, prefrontal cortex and several regions in the parietal lobule. These brain regions have consistently been associated with cognitive control processes (Carter et al., 1998; Kerns et al., 2004; Levens & Phelps, 2010; Miller & Cohen, 2001) and the experience of negative emotions (Kober et al., 2008; Phan, Wager, Taylor, & Liberzon, 2002). Lie-telling involves many different processes, such as the violation of social norms, which may come unexpected and may evoke negative reactions. This may explain why a broad network of brain regions became active when evaluating lies.

Our findings are also in line with studies on moral judgments. These studies often report activation in regions associated with cognitive (the dlPFC and dorsal ACC) and emotional (anterior insula) processes during moral judgments (Greene & Haidt, 2002; Moll et al., 2002). Our findings concur with this literature given that these regions were also more active during the evaluations of lies. Finally, prior studies on the evaluation of lies showed activation in regions of the parietal lobule (Hayashi et al., 2014; Wu et al., 2011). An important future direction is to test how this broader network of brain regions is involved in the evaluation of specific processes involved in lying. The current study made first steps in this direction by testing the role of justifications, incentive context and deviances from the norm. These findings are discussed in more detail below.

One aim of the study was to investigate the role of justifications of lies. Interestingly, the neuroimaging results confirmed that the dorsal ACC and dlPFC were active when evaluating lies, and this was more so when the lies were justifiable, which was in line with our predictions. These findings can be interpreted in terms of a potential larger conflict due to the competition between evaluating the lies as honest or dishonest. Prior studies have shown that the dorsal ACC is an important conflict monitoring region (Botvinick, 2007; Botvinick et al., 2004; Van Veen & Carter, 2006; Van Veen et al., 2009), suggesting that lies that can be justified possibly create more internal cognitive conflict compared to lies that cannot be justified.

To test this neural pattern in more detail, we evaluated the effects of several different task manipulations. First, the incentive condition, which manipulated whether individuals lied to avoid punishment or to obtain gains, did not significantly affect neural activity. Second, a possible reason why people feel more conflicted about justifiable compared to unjustifiable lies is the feeling that it is legitimate to shuffle some facts to obtain preferable outcomes (i.e., report an observed outcome on one of the irrelevant-for-pay die rolls), but not to invent facts (i.e., report a number one never observed; Shalvi et al., 2012). Indeed, the manipulations of deviance in this study showed that the deviance of the lie from the truth determined to what extent dishonest reports were considered to be lies. In addition, for these large deviance lies there was considerably larger activation in the dorsal ACC and dlPFC when participants evaluated justifiable vs. unjustifiable lies. These effects were, however, found in a contrast with an uncorrected threshold, and did not hold with a corrected threshold. Moreover, the between-subjects design for incentive frames was possibly underpowered to detect differences. Future studies could investigate the incentive and deviance hypotheses with a larger sample size.

In a way, one may argue that our study posed an odd question to participants: 'to what extent a given outcome combination is a lie'? That is, the answer to this question is categorical: either a lie or not a lie. It is not expected, however, that individuals think about dishonesty in categorical terms (Shalvi et al., 2011). As our results show, participants rated dishonesty as a spectrum ranging between absolute honesty and absolute dishonesty. This is evident from both the differences in judgment between justifiable and unjustifiable lies, as well as differences in evaluation of lies with a different deviance.

One alternative explanation for our results could be that our participants evaluated justifiable lies more leniently, because they thought the other people made a mistake in reporting the correct (i.e., first) die roll. Two findings speak against this explanation. First, our results show that there is an effect of the deviance of the lie. If justifiable lies were indeed perceived to be mistakes in reporting the correct die, then one would not expect an effect of the deviance of the lie for justifiable lies. Second, it is unlikely that the other people made mistakes in reporting the correct outcome in 20 out of 60 trials. If participants did perceive these reports as mistakes for the first part of the trials, one would expect differences between the evaluations of justifiable lies in the first and the second block of trials. We did not find any differences between the evaluations of both blocks.

Taken together, this fMRI study revealed that evaluating lies of others activates a large network of brain regions associated with social cognition. We found that evaluating lies is associated with activation in a broad network of regions associated with cognitive processes (such as the dLPFC and dorsal ACC) and emotional processes (such as the anterior insula). Moreover, we showed that lies that can be justified are rated more leniently compared to lies that cannot be justified. Furthermore, justifiable lies are considered as less dishonest when they result in a small deviance from the truth compared to a large deviance. Even though justifiable lies are rated as less dishonest, they result in larger activity in the dorsal ACC and dlPFC, two regions of the brain that are often associated with cognitive control and conflict processing. Corroborating earlier findings, our results demonstrate that justifiable lies are not considered to be complete lies. We provide neural evidence showing that the neural correlates of justifiable lies are different from the neural correlates of lies that cannot be justified, which suggests that not all lies are treated equally.

Acknowledgements

We thank Hannah Nohlen and Andreas Burger for their help with data collection.

References

Ariely, D. (2012). *The honest truth about dishonesty: how we lie to everyone, especially ourselves*. New York, NY: HarperCollins.

Ayal, S., & Gino, F. (2011). Honest rationales for dishonest behavior. In M. Mikulincer, & P. R. Shaver (Eds.), *The social psychology of morality: exploring the causes of good and evil*. Washington, DC: American Psychological Association.

Botvinick, M. M. (2007). Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. *Cognitive Affective and Behavioral Neuroscience*, *7*, 356–366.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, *8*, 539–546.

Brett, M., Anton, J. L., Valabregue, R., & Poline, J. B. (2002). Region of interest analysis using an SPM toolbox. *NeuroImage*, *16*, 497.

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

Carter, C. S., & Van Veen, V. (2007). Anterior cingulate cortex and conflict detection: uan update of theory and data. *Cognitive Affective and Behavioral Neuroscience*, *7*, 367–379.

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*, 979–995.

Fischbacher, U., & Follmi-Heusi, F. (2008). Lies in disguise: an experimental study on cheating. *Journal of the European Economic Association*, *11*, 525–547.

Gino, F., & Ariely, D. (2012). The dark side of creativity: original thinkers can be more dishonest. *Journal of Personality and Social Psychology*, *102*, 445–459.

Gino, F., Ayal, S., & Ariely, D. (2013). Self-serving altruism?: the lure of unethical action that benefits others. *Journal of Economic Behavior and Organization*, *93*, 285–292.

Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*, 517–523.

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*, 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*, 2105–2108.

Hayashi, A., Abe, N., Fujii, T., Ayahito, A., Ueno, A., Koseki, Y., et al. (2014). Dissociable neural systems for moral judgment of anti- and pro-social lying. *Brian Research*, *1556*, 46–56.

Kern, M. C., & Chugh, D. (2009). Bounded ethicality: the perils of loss framing. *Psychological Science*, *20*, 378–384.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, *303*, 1023–1026.

Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., & Wager, T. D. (2008). Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *NeuroImage*, *42*, 998–1031.

Levens, S. M., & Phelps, E. A. (2010). Insula and orbital frontal cortex activity underlying emotion interference resolution in working memory. *Journal of Cognitive Neuroscience*, *22*, 2790–2803.

Levine, T. R. (2010). A few transparent liars: explaining 54% accuracy in deception detection experiments. In C. Salmon (Ed.), *Communication yearbook 34* (pp. 40–61). New York, NY: Routledge.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review*, *1*, 476–490.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: a theory of self-concept maintenance. *Journal of Marketing Research*, *45*, 633–644.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Moll, J., Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiuolo, P. A., et al. (2002). The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience*, *22*, 2730–2736.

Pfister, R., & Janczyk, M. (2013). Confidence intervals for two sample means: calculation, interpretation:and a few simple rules. *Advances in Cognitive Psychology*, *9*, 74–80.

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*, 331–348.

Rosenfeld, J. P., Ben-Shakar, G., & Ganis, G. (2016). Detection of concealed stored memories with psychophysiological and neuroimaging methods. In W. Sinnott-Armstrong, F. Schauer, & L. Nadel (Eds.), *Neuroscience, philosophy and law*. Oxford University Press [in press]

Schweitzer, M. E., & Hsee, C. K. (2002). Stretching the truth: elastic justification and motivated communication of uncertain information. *The Journal of Risk and Uncertainty*, *25*, 185–201.

Shalvi, S., Dana, J., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Justifiable ethicality: observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, *115*, 181–190.

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

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: doing wrong and feeling moral. *Current Directions in Psychological Science*, *24*, 125–130.

Shalvi, S., & Leiser, D. (2013). Moral firmness. *Journal of Economic Behavior and Organization*, *93*, 400–407.

Somerville, L. H., Heatherton, T. F., & Kelley, W. M. (2006). Anterior cingulate cortex responds differentially to expectancy violation and social rejection. *Nature Neuroscience*, *9*, 1007–1008.

Toure-Tillery, M., & Fishbach, A. (2012). The end justifies the means: but only in the middle. *Journal of Experimental Psychology: General*, *141*, 570–583.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Van Veen, V., & Carter, C. S. (2006). Conflict and cognitive control in the brain. *Current Directions in Psychological Science*, *15*, 237–240.

Van Veen, V., Krug, M. K., Schooler, J. W., & Carter, C. S. (2009). Neural activity predicts attitude change in cognitive dissonance. *Nature Neuroscience*, *12*, 1469–1475.

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.