communications biology

ARTICLE

https://doi.org/10.1038/s42003-023-04827-w

OPEN

Reputation risk during dishonest social decision-making modulates anterior insular and cingulate cortex activity and connectivity

Lennie Dupont[1,2]✉, Valerio Santangelo[2,3], Ruben T. Azevedo[4], Maria Serena Panasiti[1,2] & Salvatore Maria Aglioti[1,2]✉

To explore the neural underpinnings of (dis)honest decision making under quasi-ecological conditions, we used an fMRI adapted version of a card game in which deceptive or truthful decisions are made to an opponent, with or without the risk of getting caught by them. Dishonest decisions were associated to increased activity in a cortico-subcortical circuit including the bilateral anterior cingulate cortex (ACC), anterior insula (AI), left dorsolateral prefrontal cortex, supplementary motor area, and right caudate. Crucially, deceptive immoral decisions under reputation risk enhanced activity of – and functional connectivity between – the bilateral ACC and left AI, suggesting the need for heightened emotional processing and cognitive control when making immoral decisions under reputation risk. Tellingly, more manipulative individuals required less involvement of the ACC during risky self-gain lies but more involvement during other-gain truths, pointing to the need of cognitive control only when going against one's own moral code.

[1] Department of Psychology, Sapienza University of Rome and CLN2S@Sapienza, Italian Institute of Technology, Rome, Italy. [2] IRCCS Fondazione Santa Lucia, Rome, Italy. [3] Department of Philosophy, Social Sciences & Education, University of Perugia, Perugia, Italy. [4] Keynes College, School of Psychology, University of Kent, Canterbury, Kent, UK. ✉email: lenniedupont@hotmail.com; salvatoremaria.aglioti@uniroma1.it

Although harmful to interpersonal interactions in financial, political, legal, and daily-life contexts, dishonesty remains ubiquitous. Classically investigated by economic, sociological, and psychological sciences, only in the last decades has dishonesty attracted the interest of the neurosciences. Specifically, an increasing number of functional neuroanatomy studies have tried to untangle the complex mechanisms behind dishonest decision-making[1–17]. Activation likelihood estimation (ALE) meta-analyses on the neural correlates of deceptive vs. non-deceptive behavior[18–20] show the involvement of a large cortico-subcortical neural network associated with complex functions related to deception. Telling a lie usually involves several mental operations that are cognitively demanding, e.g., deciding to lie, withholding the truth, monitoring whether the receiver believes the lie and so forth[19]. In keeping with this notion, neuroimaging studies have shown that deception elicits activity mainly in anterior regions involved in executive functioning (e.g. the dorsolateral prefrontal cortex, dlPFC, the ventromedial prefrontal cortex, vmPFC, the anterior cingulate cortex, ACC), in emotional and interoceptive processing (e.g. the anterior insula, AI), and in reward processing and inhibitory control (e.g. the nucleus caudate, Cau)[18–20]. The involvement of posterior regions like the inferior parietal lobule (IPL) involved in inhibition and selective attention and the temporoparietal junction (TPJ) involved in Theory of Mind processes has also been reported[18–20]. However, given the complexity of deception, a variety of personality, cognitive, and emotional factors seem to orchestrate controlled and automatic decision-making across individuals, with some being highly susceptible to the temptation to lie (Will hypothesis) and others being 'naturally' immune to moral violations (Grace hypothesis)[12]. Importantly, there is evidence to suggest that cognitive control may allow cheaters to behave honestly and honest people cheat depending on the circumstances[21].

It is worth noting that the initial exploration of the neural correlates of dishonest decision-making was based on tasks where the experimenter specifically instructed participants when to lie and when to tell the truth[1,14,17,22–24]. Thus, lack of intentionality and absence of social contexts made dishonest decision-making very different from what happens in real-life conditions[25]. Aware of the need to explore dishonesty in improved contexts[26], subsequent neuroimaging studies have used paradigms devised to circumvent the oversimplified laboratory conditions that are unfit to capture the complexity of (dis)honest decision-making[12,25,27]. Yet, most of the existing studies missed one or more of the features that make a laboratory paradigm as ecological as possible, namely, i) intention to lie, i.e., when the choice to lie is spontaneous instead of being instructed; ii) social interaction, i.e., when the deception occurs in a social context; and iii) motivation, i.e., when telling a lie entails a benefit or avoidance of a penalty for the liar[25]. While previous important studies (e.g. Garrett et al., 2016; Baumgartner et al., 2009)[27,28] took into account the role of crucial features in modulating deceptive behavior, in the present study we tested the effect of both dispositional (e.g. Machiavellian intelligence) and situational variables (e.g. reputation risk) in a within-subject design. To explore the pattern of neural activity during dishonesty in quasi-ecological conditions, we combined fMRI with a new version of a behavioral task, the temptation to lie card game (TLCG), that we developed in previous studies and that proved adept to tap multiple facets of spontaneous social deception[29–34].

The TLCG is an interactive card game where experimental participants observe another player (who unbeknownst to the participant is a computerized opponent) choosing one out of two covered cards, namely, the ace of hearts or spades. Picking one or the other implied winning or losing money, respectively, from a common pocket. The participant is the only one who can see the choice outcome (which can be favorable or unfavorable depending on whether they are supposed to win or lose) and has the liberty to accept or reverse the outcome. For example, when the outcome implies a loss for the participant they can report a win instead, thus making a dishonest decision (self-gain lie). Therefore, the TLCG includes intentionality, by letting the participants make spontaneous decisions to lie or tell the truth; sociality, by including an opponent whom the participant played against; and motivation, by including a monetary reward when they win on a trial-by-trial basis. In addition, an important and somewhat neglected facet of deception explored by our approach is the probability of getting caught in one's lies. Indeed, our participants were informed that in some trials nobody could see their decision (no-reputation risk) while in others the opponent could be informed about their decision (reputation risk). We refer to the last condition as 'reputation risk' due to the risk of losing the social prestige, or capital, associated with being regarded as trustworthy. Thus, any immoral decision jeopardized the participant's moral reputation in the latter but not in the former type of trial. Of note, the loss of reputation has been shown to be so relevant that people may prefer undergoing very unpleasant experiences, such as physical pain, to avoid it[35]. One thermal imaging study in healthy individuals indicated that self-gain lies under reputation risk conditions influenced variations of nasal temperature that can be ascribed to para-sympathetic activity. This result suggests that the risk of being caught in a lie is associated with the need to regulate one's own emotional activity[29]. Moreover, one study indicates that poor interoceptive accuracy leads to a stronger effect of the reputation risk on the reduction of lies[33], suggesting that being less aware of one' own bodily signals makes us more susceptible to the effect of social contexts.

The main aim of our study was to determine the neural network involved in spontaneous dishonest decision-making in general and in the processing of lies, as well as its modulation by the risk of getting caught by the opponent and thus losing one's own reputation. Moreover, we wanted to explore whether individual differences in morality were associated with the modulation of dishonest decision contingent upon reputation risk. Based on the above-mentioned studies, we hypothesized that reputation risk will reduce dishonesty. However, when making a risky decision for one's own reputation, the involvement of the executive control regions already associated with dishonesty might increase compared to anonymous dishonest decisions, especially for individuals who consider themselves more honest.

This study found increased neural activity within a circuit encompassing the bilateral ACC, AI, left dlPFC, supplementary motor area (SMA), and right Cau during dishonest decisions. Furthermore, the results show that when individuals make dishonest decisions for a monetary reward under reputation risk, the bilateral ACC and AI exhibit increased activity and connectivity, indicating the necessity for enhanced emotional processing and cognitive regulation. Notably, individuals with higher levels of manipulativeness were found to require less ACC involvement while deceiving for their own benefit, but more involvement when telling the truth for others' advantage. This finding suggests that cognitive control is only necessary when acting against one's own moral values. In conclusion, lying may require varying degrees of cognitive effort, depending on social risk factors and on individual's dispositional traits.

Results

Behavioral results. We tested whether participants' lying behavior (both for self-gain and other-gain motivations) was influenced by the fact that they were risking being caught by their opponent. Firstly, we tested whether the outcome (Fav or UnFav) and risk (No Rep or Rep) influenced lying percentages and response times. Secondly, we wanted to know whether certain

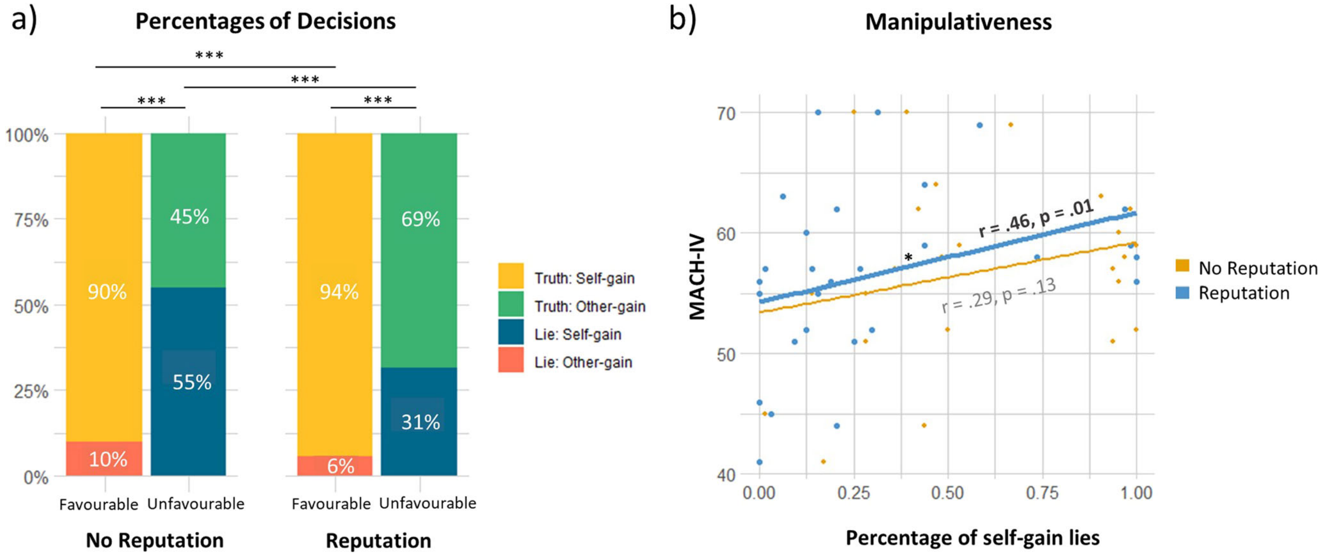


Fig. 1 Decision-making percentages and correlations with levels of manipulative traits. a Percentages of decisions for each reputation condition (No Rep or Rep) and within each outcome (Fav or UnFav) ($N = 28$). **b** Correlation graph of individuals' Machiavellian scores with the percentages of self-gain lies during no reputation risk (yellow) and reputation risk (blue) conditions ($N = 28$). The line plot in bold depicts significance. MACH-IV Machiavellian questionnaire scores.

dispositional personality factors predicted the probability of lying for both the no reputation risk and reputation risk condition. To deal with this question, we correlated questionnaire scores with other-gain (OG-Lies) and self-gain lying (SG-Lies) percentages in both conditions.

In agreement with previous literature[30], the results of the binomial 2 × 2 generalized linear mixed model showed that the lying percentage was modulated by the interaction between outcome and reputation (Fig. 1a, Interaction effect: Lie ~ Out × Rep: $\chi^2(1) = 7.49$, $p = 0.006$). As expected, when the outcome was unfavorable, participants lied significantly more for self-gain when they were guaranteed anonymity compared to when their reputation was at risk (z-ratio = 4.55, $p < 0.0001$). Moreover, fewer OG-Lies were made in the reputation condition compared to the no reputation condition (z-ratio = 3.14, $p = 0.009$). OG-Lies, however, were made much less often compared to SG-Lies in both conditions (No Rep: z-ratio = −6.16, $p < 0.0003$, Rep: z-ratio = −5.82, $p < 0.0001$).

Moreover, we found that manipulativeness scores correlated positively with SG-Lie percentage during reputation risk (Fig. 1b, $r = 0.46$, $p = 0.01$), while this was not significant for the no reputation condition ($r = 0.29$, $p = 0.13$). Thus, the more subjects were manipulative the more they tended to make SG-Lies in the reputation condition, while this positive correlation was not significant when they were anonymous. No other correlations were found between the lie probability and the Balanced Inventory of Desirable Responding (BIDR) and the Civic Moral Disengangemet (CMD).

Lastly, we find that truthful responses take significantly less time than lies [$\chi^2(1) = 20.69$, $p < 0.0001$] and decisions are made faster when the outcome was favorable [$\chi^2(1) = 19.14$, $p < 0.0001$] (see Fig. S3). A significant interaction effect was found between all 3 factors (reputation × outcome × decision: $\chi^2(1) = 5.07$, $p = 0.02$). Post-hoc tests revealed that for both the no reputation and the reputation condition, self-gain truths take the least amount of time compared to all the other conditions.

fMRI results. We used a first general linear model (GLM1; see section "fMRI data analysis") to investigate the main effect of spontaneous lying vs. truth-telling and the interaction between spontaneous lying and the reputation conditions (Rep, vs. NoRep), independently of the outcome (i.e., Fav or UnFav). A contrast image of "all spontaneous" versus "all instructed trials" was used as a restricted search area (see Fig. S5 in Supplementary material). This analysis revealed that the main effect of spontaneous lying (vs. truth-telling) recruited a brain circuit involving anterior regions, including the ACC, AI, bilaterally, the left dlPFC, the right supplementary motor area (SMA), and the right caudate (Cau) (Fig. 2a, purple maps and Table 1). As indicated by the related signal plots (Fig. 2b, gray bars), only the right Cau showed an increase of activity for lie conditions irrespective of reputation risk (compare bars 1 & 3 vs. bars 2 & 4), while for the other regions, the activity appeared to increase when participants lied under reputation risk as compared to the other conditions, i.e., an interaction pattern (compare bar 3 vs. the other bars in the signal plots of Fig. 2b). However, a significant interaction effect of reputation (Rep, NoRep)×decision (Lie, Truth) was formally observed only in one of these regions, namely, in the ventromedial portion of the left anterior insula (vmAI; yellow bar plot in Fig. 2b and Table 1). The vmAI showed a selective increase of activity following the Rep_Lie condition. No other interaction effects were found within this model.

A second GLM analysis (GLM2) allowed us to further investigate spontaneous lying under reputation risk, highlighting the differences between self-gain and other-gain motivations: i.e., SG-Lies and OG-Truths (unfavorable outcome); SG-Truths and OG-Lies (favorable outcome). We here focused on decisions made when the outcome was unfavorable because this outcome is arguably the most interesting condition as it contrasts SG-Lie and OG-Truth, which we are most interested in. A contrast image of all spontaneous versus all instructed trials was used as a restricted search area (see Fig. S5).

GLM2 revealed a main effect of SG-Lies vs. OG-Truths in the right ACC (Fig. 3a, red region & Table 2a), although the activity in this region appears to be modulated by the reputation condition (Fig. 3b, red bars 1 & 4 vs. red bars 2 & 5). In fact, the right ACC, along with the left ACC and the left vmAI, showed an interaction effect for SG-Lies vs. OG- Truths with reputation risk (Fig. 3a, blue regions and Fig. 3b, blue bar plots, & Table 2b). This

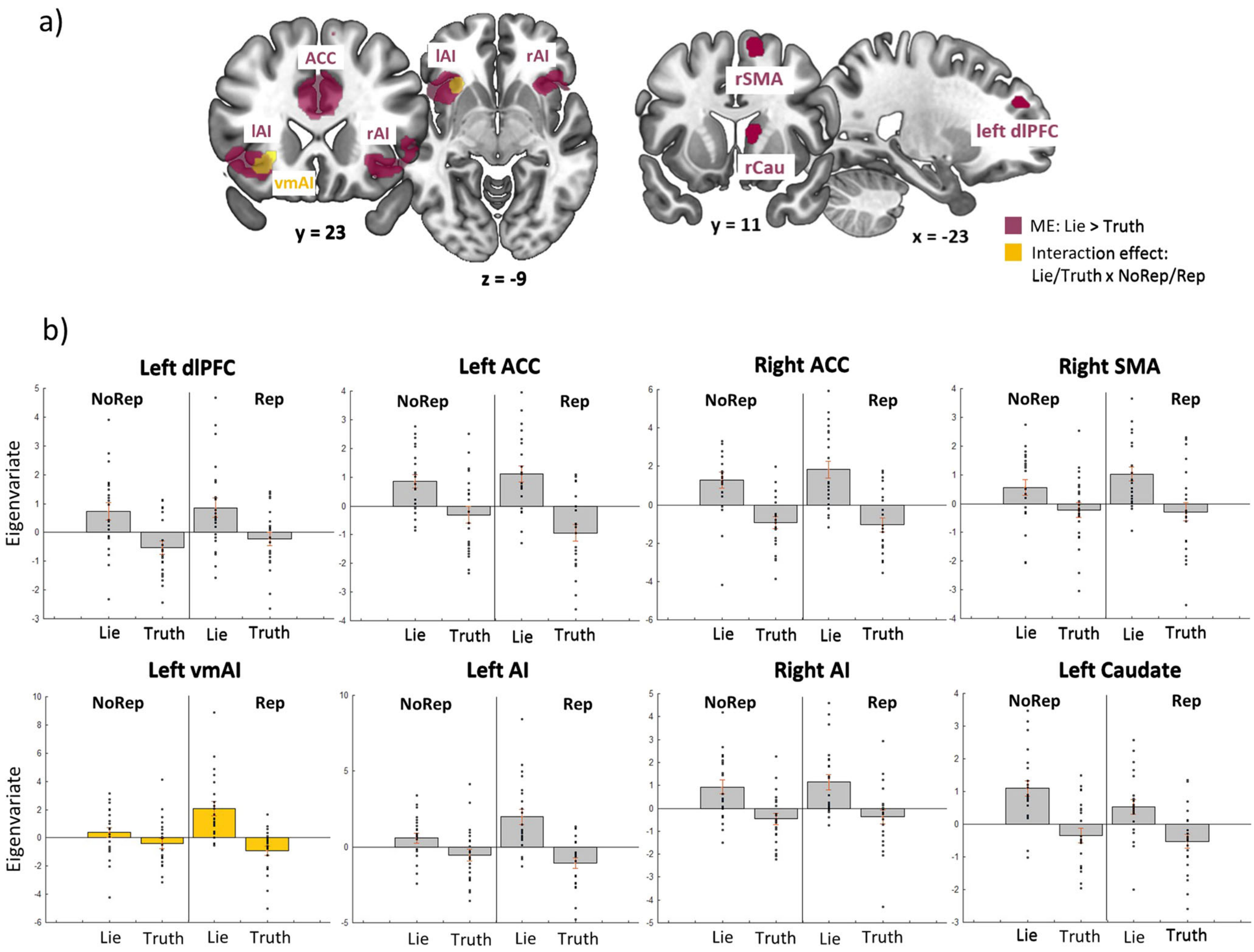


Fig. 2 Brain activation related to spontaneous lies in general and lies under reputation risk. Results of GLM1: **a** coronal and axial slices showing the regions active for lies vs. truths in both conditions, i.e., ACC and AI (purple map) and the left ventromedial section of the left AI for lies in the reputation risky condition (yellow map) ($N = 22$). **b** Bar plots of regions more active for lies vs. truths. Gray bar plots for lying in both conditions, and yellow bar plots for spontaneous lies in the reputation risk condition ($N = 22$). ACC anterior cingulate cortex, lAI/rAI left/right anterior insula, rSMA right supplementary motor area, dlPFC dorsolateral prefrontal cortex, rCau right caudate, NoRep no reputation risk, Rep reputation risk, BOLD blood-oxygen-level-dependent response.

Table 1 Brain activation for the main effect of "spontaneous lies vs. truths" (NoRep_Lie + Rep_Lie > NoRep_Truth + Rep_Truth) and for the interaction (Rep_Lie + NoRep_Truth > NoRep_Lie + Rep_Truth).

L/R	Regions	kE	x	y	z	Z	P-FWE-corrected
Main effect of Lie							
R	ACC	1835	4	28	28	5.69	<0.001
R	ACC		4	36	28	5.66	<0.001
L	ACC		−6	28	28	5.32	<0.001
L	ACC		−10	30	20	5.13	<0.001
R	ACC		6	44	20	4.75	0.002
L	ACC		0	48	18	4.7	0.002
L	AI	237	−34	22	−8	5.52	<0.001
L	AI		−42	16	−8	5.01	<0.001
R	AI	99	30	18	−10	4.76	0.002
R	SMA	158	10	14	62	4.19	0.018
L	dlPFC	23	−22	46	30	4.09	0.026
R	CAU	19	12	14	12	4.02	0.034
Interaction: Rep - No Rep × Lie - Truth							
L	AI	60	−32	22	−10	4.04	0.031

Only clusters and their local maxima that survived family-wise error (FWE) correction ($p < 0.05$) are presented.
L/R left/right, *kE* number of voxels, *p-FWE* family-wise error corrected *p*-value, *ACC* anterior cingulate cortex, *AI* anterior insula, *SMA* supplementary motor area, *dlPFC* dorsolateral prefrontal cortex, *CAU* caudate.

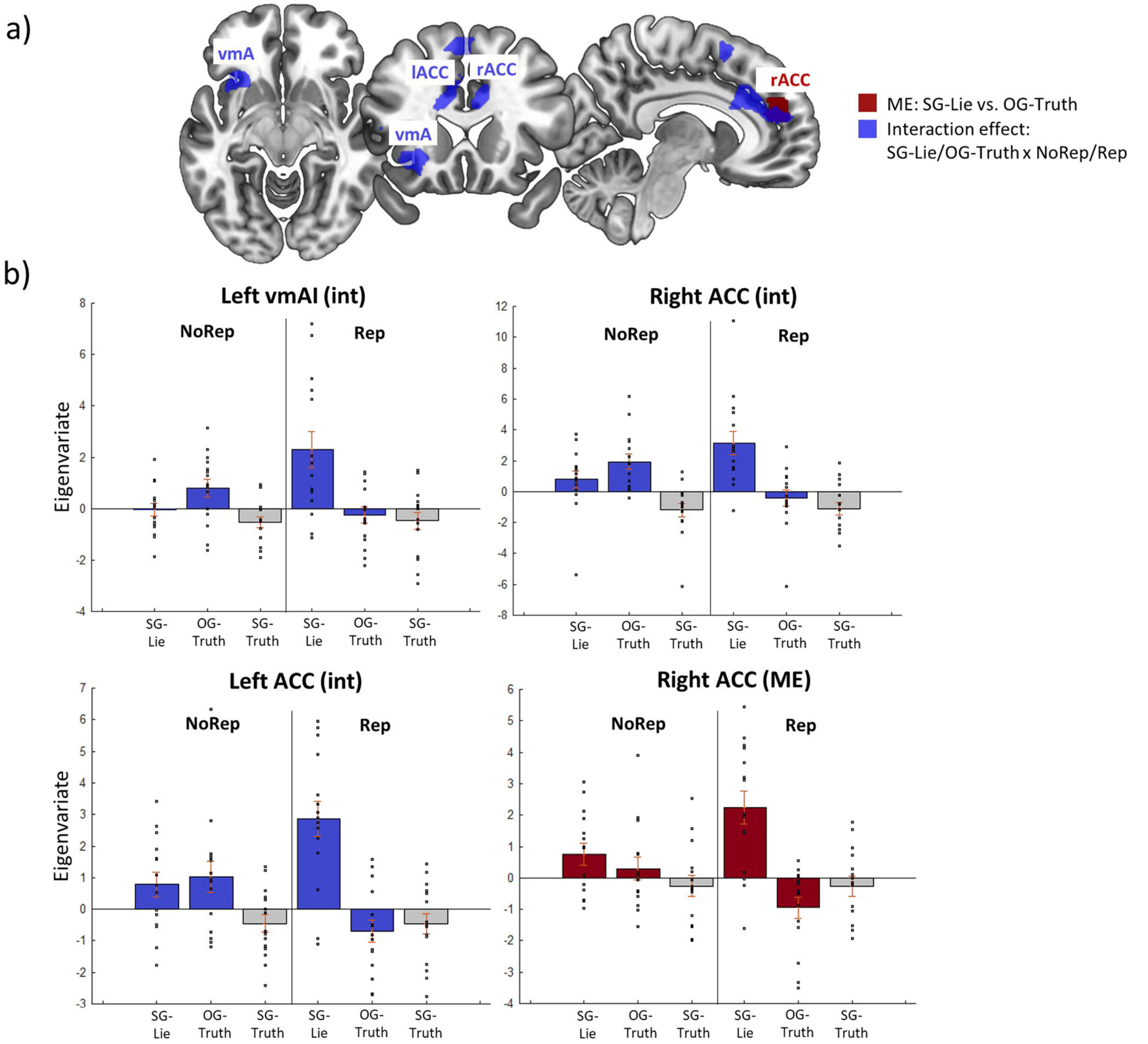


Fig. 3 Brain activation related to self-gain lies compared to other-gain truths and self-gain lies under reputation risk. Results of GLM2: **a** axial, coronal, and sagittal slices showing, in red, the regions active for self-gain lies vs. other-gain truths irrespective of reputation conditions and, in blue, the regions that are most active for self-gain lies vs. other-gain truths with reputation risk, namely, the left vmAI, and the bilateral ACC (interaction effect: self-gain lies/other-gain truths × reputation/no reputation) ($N = 15$). **b** Bar plots showing, in blue, the increased activation of the left vmAI and left and right ACC for self-gain lies with reputation risk (bar 4) and, in red, showing increased activation for self-gain lies (bars 1 & 4) vs. other-gain truths (bars 2 & 5) in both reputation conditions ($N = 15$). vmAI ventromedial anterior insula, lACC/rACC left/right anterior cingulate cortex, ME main effect, int interaction effect, SG-Lie self-gain lies, OG-Tr other-gain truth, SG-Tr self-gain truth.

Table 2 Brain activation for the contrasts: (a) NoRep_SG-Lie + Rep_SG-Lie > NoRep_OG-Truth + Rep_OG-Truth and (b) Rep_SG-Lie + NoRep_OG-Truth > NoRep_SG-Lie + Rep_OG-Truth.

L/R	Regions	kE	x	y	z	Z	P-FWE-corrected
(a) Main effect: Self-gain Lie > Other-gain Truth							
R	ACC	694	8	48	20	4.6	0.009
(b) Interaction: Rep/No Rep × SG-Lie/OG-Truth							
L	vmAI	225	−28	18	−6	4.42	0.018
R	ACC	1566	8	28	30	4.21	0.041
L	ACC		−8	40	16	4.18	0.046
R	ACC		8	32	26	4.18	0.047

Only clusters and their local maxima that survived family-wise error (FWE) correction ($p < 0.05$) are presented.
L/R left/right, *kE* number of voxels, *p-FWE* family-wise error corrected *p*-value, *ACC* anterior cingulate cortex, *vmAI* ventromedial anterior insula.

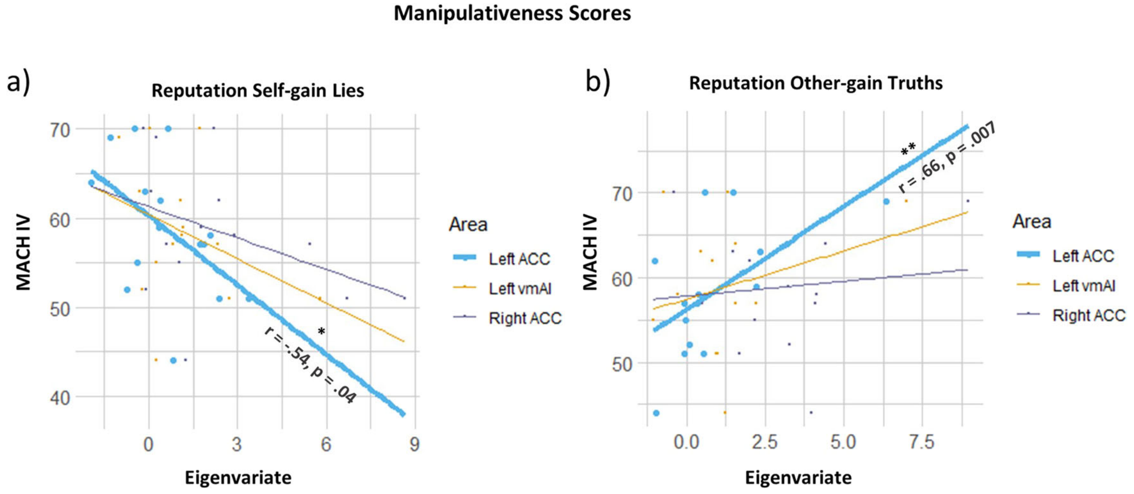


Fig. 4 Correlations of levels of manipulative traits and decision-related brain activity in the regions of interest. a Correlation of averaged beta-values in three brain areas (LACC, LvmAI, RACC) during self-gain lies with reputation risk and Machiavellian scores. **b** Correlation of averaged beta-values in three brain areas (LACC, LvmAI, RACC) during other-gain truths with reputation risk and Machiavellian scores. The line plots in bold depict significance ($N = 15$). ACC anterior cingulate cortex, vmAI ventromedial anterior insula.

means that these regions are selectively more active for SG-Lies vs. OG-Truths under reputation risk as compared to no-risk conditions (compare bars 4 & 5 vs. bars 1 & 2 in Fig. 3b).

As these regions are elicited the most for SG-Lies under reputation risk (i.e., the main aim of the current study), we used them as our main regions of interest (ROIs) for the following correlation and connectivity analyses.

The main effect of lies (self-gain) vs. truths (self-gain and other-gain) elicited similar regions as the main effect of GLM1 (see Fig. S2A, green map, and Table S4.1a) and the same interaction was found in the left ventromedial insula with reputation risk (see Fig. S2A (blue) and Table S4.1b). When contrasting SG-Lies with SG-Truths for both conditions bilateral ACC and AI activation is found, together with right SMA (see Fig. S2B, yellow map, and Table S4.2a). Finally, contrasting OG-Truths with SG-Truths for both conditions elicits bilateral ACC and right SMA (see Fig. S2.C, orange map, and Table S4.3a). No interaction effect with reputation was found for the final two contrasts.

Correlation analyses. We tested whether the BOLD signal of the main regions of interest found in the interaction effect (i.e., the left vmAI, and the left and right ACC) correlated with qualitative measures. We found an inverse relationship between MACH-IV scores and BOLD activity in the left ACC during self-gain lies under risk, i.e., the higher participants scored on manipulativeness the lower the activity in the left ACC for self-gain lies under the reputation risk condition (Fig. 4a, $r = 0.54$, $p = 0.04$). Contrarily, the left ACC was less active during other-gain truth under risk for less manipulative participants (Fig. 4b, $r = -0.66$, $p = 0.007$). Less manipulativeness was also associated with increased activity in the left vmAI during NoRep_SG-Lies (see Fig. S4, $r = -0.60$, $p = 01$).

Connectivity analyses. Finally, we explored whether and how the pattern of functional connectivity changed for the three ROIs derived from the interaction effect of SG-Lies vs. OG-Truths (i.e., the left vmAI, and the left and right ACC) as a function of reputation risk. Relative to other-gain truth-telling, self-gain lying with reputation risk induced a stronger coupling of the left vmAI with the bilateral ACC: right ACC: $F(2,13) = 7.59$, $p = 0.006$, left ACC: $F(2,13) = 4.82$, $p = 0.02$ (Fig. 5a and b, yellow bar vs. purple bar). In the no reputation risk condition, this relationship is reversed (see Fig. 5a and b, red bar vs. blue bar). The same contrast showed, however, the strongest connectivity between the right ACC and the left dmPFC for Rep_SG-Lie [$F(2,13) = 4.22$, $p = 0.03$] (Fig. 5c, yellow bar).

Discussion

Using an ecological approach wherein participants could freely decide to lie or to tell the truth, we examined the impact that reputation risk and game outcome factors exert on spontaneous dishonest social decision-making, both behaviorally and in terms of neural activity modulations. We have been able to determine the regions more active when spontaneous lies vs. truths were made independently from reputation condition and motivation of decisions (self-gain or other-gain). Notably, we highlighted the brain areas recruited more for lies in general and self-gain lies in reputation risk conditions and found high vs low ACC activity in high vs low manipulativeness. Finally, we were able to determine the strength of the connections between the nodes in the neural circuit underpinning the deceptive and truthful decisions in our ecological task.

Behavioral findings. We found that both reputation risk and outcome factors influenced participants' dishonest decisions. In specific, risking one's own reputation reduces self-gain dishonesty and unfavorable game outcome increases it. These behavioral findings are in line with those reported by Panasiti and colleagues[29–31] and Azevedo and colleagues[32]. It is worth noting that a high occurrence of self-gain lies has been found in previous studies based on a similar version of our experimental paradigm[29–31,33,34,36]. The finding that other-gain lies are higher in the no reputation risk may be linked to the human tendency of rather donating anonymously than with a chance of being found out[37]. That this result was not found in a previous study where we used a similar protocol[30] may be due to the difference in effect size of the reputation condition on lying behavior. Moreover, we found that lying takes longer than truth-telling, corroborating the previous findings[38–40]. However, there was no significant difference in reaction times between outcome (unfavorable vs. favorable) and reputation risk conditions.

Brain regions involved in lying compared to truth-telling. When exploring the brain regions associated with lying relative to

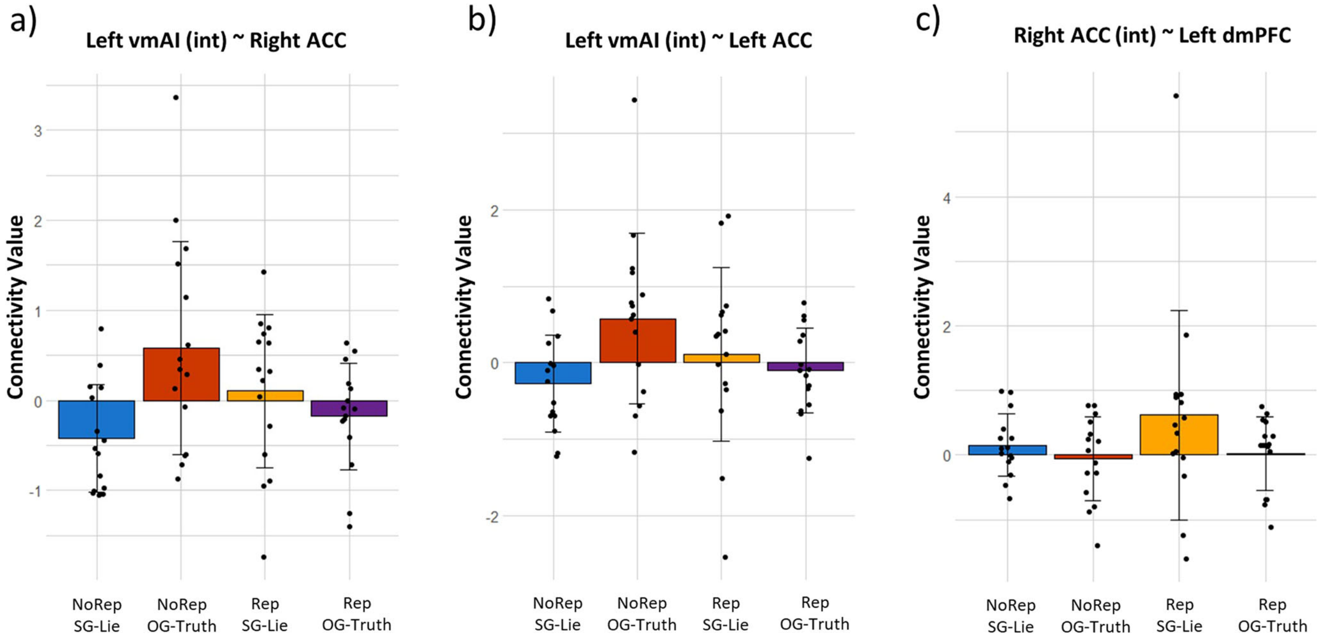


Fig. 5 Connectivity analysis results. a Stronger connectivity between the left vmAI and right ACC for self-gain lies with reputation vs. no reputation risks (yellow vs. blue bar) and for other-gain truths with no reputation vs. reputation risk (red vs. purple bar). **b** Stronger connectivity between the left vmAI and left ACC for self-gain lies with reputation vs. no reputation risks (yellow vs. blue bar) and for other-gain truths with no reputation vs. reputation risk (red vs. purple bar). **c** Stronger connectivity between the right ACC and left dmPFC for self-gain lies with reputation risk vs. other conditions (yellow vs. blue, red, and purple bars). vmAI ventromedial anterior insula, ACC anterior cingulate cortex, dmPFC dorsomedial prefrontal cortex, NoRep_SG-Lie self-gain lies without reputation risk, NoRep_OG-Truth other-gain truths without reputation risk, Rep_SG-Lie self-gain lies with reputation risk, Rep_OG-Truth other-gain truths with reputation risk.

truth-telling regardless of the reputation risk and outcome factors, we found changes of activity in a neural circuit, including the bilateral ACC and AI, the left dlPFC, the SMA, and the right Cau. The activation of these anterior regions of the brain is in good agreement with meta-analyses on the neural correlates of dishonesty[18–20,41]. In particular, activation in the ACC and the insula have been consistently found for deceptive compared to truthful behavior across a variety of tasks and stimuli. The ACC has been linked to a wide range of executive control processes that are needed to execute a deceptive response, such as social context integration[42], working memory, inhibition of competing responses, mediation of cognitive conflict[39,43], task-switching, reward, interoception and motivation in decision-making[2,20,44–47]. The bilateral AI, besides its general involvement in executive control, has also been linked to the visceral responses (e.g., blood pressure, heart rate, body temperature) and interoceptive activity[48] that accompany deceptive behavior[19,20]. The ACC and AI are typically considered crucial nodes of the Salience Network (SN), whose activity has been linked to the processing of behaviorally salient events[49]. The engagement of the SN may play a crucial role in cognitive control[50]. This supports the idea that the AI functions as an integrative center that identifies salient events from lower-level sensory inputs and relays this information to the task- and context-relevant brain networks to facilitate attention allocation, working memory[50], and decision-making operations[51,52]. Moreover, the dlPFC has been linked to spontaneous vs. instructed dishonesty[53] and has been found to play a role in executive processes regardless of the deception task used[22].

It is noteworthy that most of the brain regions activated for deception exhibited heightened activity when lies are produced in reputation risk conditions (see Fig. 2b). This, however, did not apply to the right caudate, which appears to be equally or even slightly more activated during anonymous lying. The caudate nucleus is considered a key node in the reward network[49] and its activation in response to lying might reflect the anticipation and appraisal of the financial reward associated with the act of lying[15,50]. Additionally, higher resting-state connectivity between brain networks, such as the vmPFC, associated with self-referential thinking) and the Cau (associated with reward), has been shown to correlate with an individual's tendency to cheat[15]. The equal activation of this region in both anonymous and reputation-risk conditions could be because participants would receive the same reward for lying in both situations, leading to similar reward anticipation and appraisal.

Brain regions involved in self-gain lying when reputation is at risk. Our ecological task allowed us to explore the activity of a lying-related network when the risk of being caught in a lie was or was not present. Under reputation risk, we found a selective activation of the ventromedial portion of the left anterior insula for lying (GLM1). The anterior insula is known not only for its role in interoceptive awareness and visceral responses[54], but also for its involvement in socially relevant functions like social exclusion[55], exposure to unfair treatment[56], the anticipation of reputation decisions[51], and when making inequitable decisions[57]. Studies have shown that the activation in the AI is predictive of subsequent immoral behavior[28] and its activation is associated with the anticipation of guilt, and emotion that is crucially involved in moral decision making[58,59]. Lying for altruistic compared to self-serving purposes has been found to reduce AI activity[40]. Research in neuroeconomics has shown that the processing of financial risk-taking when there are potential losses involved is mainly associated with activity in the anterior insula[56,60,61]. It is hypothesized that this is due to the processing of more aversive emotions related to risk-taking or risk anticipation[62] than control conditions. Specifically the ventral AI, compared to the dorsal AI, has been associated with more affective processes, e.g., mediating aversive feelings that generate motivation to norm enforcement[63]. Based on these findings, a plausible interpretation of the activation of the left vmAI during deception under reputation risk found in our study is likely due to heightened processing of emotions that arise with risk-taking, such as fear, sadness, disgust, anxiety[64,65], or guilt[66]. Here we further add to the literature that the left vmAI plays a role in the modulation of dishonest decision-making under reputation risk.

Diving deeper into the neural nature of decision-making under reputation risk, we found that when the outcome was unfavorable, i.e., the opponent wins and the participant's tendency to lie is maximal, there is an increase in vmAI activity together with the left and right ACC for self-gain lies with a reputation risk. This is in line with a previous study reporting that the presence of an audience reduced the likelihood of accepting an immoral offer for monetary gain and that audience vs. no audience engaged a brain network including the anterior insula, the ACC, and the right TPJ[63]. These regions may reflect meta-representations of what other people think about us or about our desire for social norm compliance[63]. We argue that going against social norms (in this case, honesty) in favor of a monetary reward, while there is a risk of getting caught, is emotionally more salient than going against social norms without social risk. This may entail increased bottom-up activity in the vmAI and, in turn, the recruitment of the ACC needed for increased executive control and mediation of cognitive conflict.

Our findings are in seeming contrast with a study reporting increased subgenual ACC activity during deceptive decisions when there was no risk of confrontation[67]. However, the risk of confrontation in this study means monetary penalization, while in ours it means loss of reputation without loss of money. This difference could make one task more sensitive than the other in grasping the role of the ACC in monitoring reputation risk.

It is also noteworthy that our functional connectivity analysis showed a stronger coupling of the left vmAI with the left and the right ACC during self-gain lies compared to other-gain truths in the reputation risk condition. Tellingly, this relationship was reversed in the no reputation condition. This pattern of results may reflect the interplay between emotional and executive processes. Cognitive conflict seems to arise both when selfish lies are made with a potential social penalty but monetary reward and when altruistic truths are told with no social reward and a monetary penalty. Similarly, the ACC was found to be more connected to the dmPFC, a region often associated with cognitive control and dishonest decision-making[12,13,21]. We suggest that this increased connectivity could be attributed to the ACC transmitting signals pertaining to the detection of conflicts, necessitating the dmPFC to provide top-down cognitive control[68,69] specifically in the context of dishonest decision-making when reputation is under threat.

Individual differences in behavioral and brain correlates of deception. Another interesting result of our study is that the manipulative characteristics of our participants correlated both with the tendency to deceive and the brain correlates associated with this behavior. Higher Machiavellian scores were associated with a higher production of self-gain lies in the risk reputation condition and with a decrement of activity in the left ACC in the very same condition. Individuals with higher MACH-IV scores typically have less problems deviating from social norms. That more manipulative individuals making risky selfish lies do not need to recruit the ACC as much may suggest that their decision to violate the norm requires less conflict monitoring and therefore less executive control. Both findings are in accordance with our previous research indicating that higher manipulative traits were associated with a smaller effect of reputation risk[30], a smaller inhibition of the cortical motor readiness to lie[31], and a smaller regulation of the sympathetic system during lying when reputation is at risk[29].

In the present study, we also found reduced left ACC activity for other-gain truth-telling under reputation risk in less manipulative individuals. In sum, highly manipulative individuals seem to need more cognitive control when making honest decisions under risk and low manipulative individuals need it when making dishonest decisions under risk. These results are in keeping with the theory[21] that cognitive control is needed to override one's own moral default[16]. However, further research is needed to assess the specific link between cognitive control and reputation risk".

Possible limitations and conclusions. It is important to acknowledge that the motivations behind dishonest behavior are complex and multifaceted. Our study focuses solely on dishonesty driven by financial self-interest and does not examine other forms of dishonesty, such as prosocial dishonesty[13,40,70], which occurs when lying is motivated by altruistic reasons, or sophisticated deceptive behavior[53,71], in which the truth is used to deceive others. The limited occurrence of prosocial dishonesty in our study may be due to the absence of incentives for this behavior. Further research, such as that conducted by Azevedo and colleagues (2018)[32], suggests that the provision of additional information about the opposing party can foster empathy and thereby increase prosocial dishonesty.

Another potential limitation is that we had unbalanced trial counts given the ecological nature of the task as the different participants made rather diverse choices. While we acknowledge that this may lead to a reduction of statistical power, we note that our approach provides a veridical picture of what happens under real-life circumstances where complex decision behaviors come with high interindividual variability at both dispositional and situational levels. On the other hand, however, our approach is reminiscent of what happens under life circumstances where complex decision behaviors come with high interindividual variability at both dispositional and situational levels. Finally, to understand the interactions and dynamics of the network found in this study, a more detailed analysis could be done in the future such as dynamic causal modeling (DCM) to gain a further understanding of the underlying mechanisms of lying under reputation risk.

Notwithstanding the current limitations, our experimental approach allowed us to reveal some of the important regions needed for making dishonest decisions when one's own reputation was or was not at risk. The recruitment and the increased functional connectivity of the anterior insular and cingulate cortex when making dishonest decisions under the influence of a social risk points towards a greater need for emotional processing and executive functioning, likely because going against one's own social norms (honesty before money) causes inner conflicts. When one's own social norms, however, shift toward selfishness (money before honesty) like, in the more manipulative individuals, the need for conflict monitoring seems to arise more with altruistic honest decisions and less with selfish dishonest decisions in a social context.

Methods

Participants. Thirty-four participants (19 females, range = 20–46 years, mean = 26.19 ± 5.65 years) enrolled in the fMRI experiment. All participants had a normal or corrected-to-normal vision, were free from any contraindication to fMRI, and had no history of major psychiatric or neurological problems. All participants gave their written informed consent, and the study was approved by the independent

Ethics Committee of the Santa Lucia Foundation IRCCS (Scientific Institute for Research Hospitalization and Health Care). Data from six participants were excluded prior to analyses (three participants were excluded due to technical problems related to the acquisition of the anatomical scan; two did not believe that the opponent was a real player; and one did not understand the task properly), leaving a final sample of 28 participants (14 females, range = 20–46 years; mean = 26.45 ± 5.95 years).

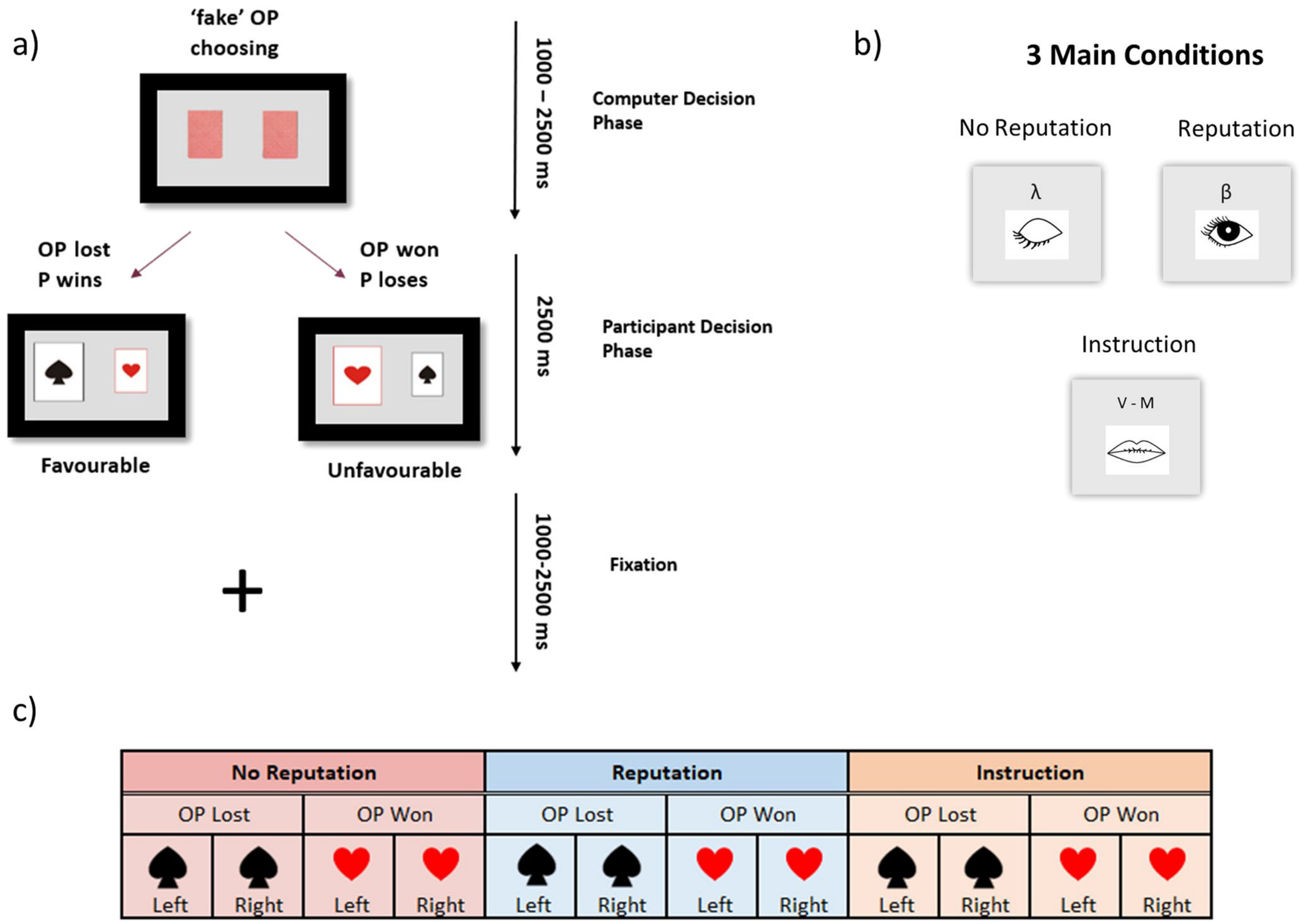
The appropriate sample size for this study was estimated with G*Power 3.1.9.2 (ANOVA, repeated measures, within factors), considering a medium effect size of 0.20 (predicted based on Panasiti et al., 2011[30], using the same design), a significance level of 0.05, 1 group, 12 measurements (i.e., 3 reputation condition × 2 decisions × 2 outcomes). This indicated a power > 95% using a sample size of 28 participants.

Task. We used the TLCG, a card game to explore dishonest decision-making in a social context. This type of paradigm proved adept to highlight situational and dispositional factors as well as social variables that may influence the participants' performance[29–32]. The TLCG was adapted for the scanner session. Inside the MRI scanner, participants played the TLCG against a computerized opponent (OP). Crucially, however, participants were told they were playing against a real opponent who was sitting in a different room and whom they would meet at the end of the experiment. Only at the end of the experiment were participants fully debriefed and told that the OP's choices were made by a computer algorithm.

Each trial of the TLCG started with the OP choosing one of two covered cards, one on the left and one on the right side of the display, within a time window of 1–2.5 s (Fig. 6a). The chosen card could be either the ace of spades or the ace of hearts, indicating a loss for the OP (favorable outcome for the participant) or a win for the OP (unfavorable outcome for the participant), respectively. Prior to the experiment, participants were informed that only they could see the outcome of the OP's choice and communicate it to the OP. Participants were free to communicate the outcome truthfully or to reverse the outcome. Decisions were communicated by pressing one of two response keys within a time window of 2.5 s. For instance, the participant could communicate that the OP chose the losing card while, in reality, the OP had chosen the winning card (i.e., Self-gain lie) or s/he could communicate that the OP chose the winning card while OP had chosen the losing card (i.e., Other-gain lie). After a jittered inter-trial interval (mean of 1.75 s, range = 1–2.5 s) filled with a fixation cross, a new trial began. Participants were aware that in each trial the monetary pay-off could go only to the winner and that a different amount of money was associated with each trial. The exact gain would only be communicated at the end of the game to rule out that participants' choice was based on a trial-by-trial computation of gain and loss and to ensure that the temptation to deceive was comparable across trials. Participants were given 10 euros for their participation and could win up to 25 euros extra during the game.

The game was performed under three main conditions (Fig. 6b), two spontaneous and one instructed: (1) No reputation risk (No Rep; indicated with an eye closed symbol and a λ, lambda), in which the participants knew their decision would be unknown to the OP; (2) Reputation risk (Rep; indicated with an eye open symbol and a β, beta), in which they knew there was a 75% chance of the OP finding out their decision; and (3) Instruction (Ins; indicated with a mouth symbol and either the letter "V" for the Italian word "Verità", truth, or the letter "M" for the Italian word "Menzogna", lie, depending on whether participants were instructed to lie or telling the truth). In this condition, participants were instructed about which specific decision they had to make.

The experiment included a total of 384 trials, given by the crossing of 2 (left vs. right card selected by OP) by 2 (favorable vs. unfavorable outcome for participants, i.e., OP chose ace of spades or ace of hearts, respectively) by 3 (No Rep, Rep, or Ins experimental conditions) by 32 repetitions of each type of trial. The total experiment consisted of 4 functional MRI runs including 96 trials each. These were administered in 9 mini-blocks, 3 for each condition (i.e., No rep, Rep, or Ins), with



No Reputation				Reputation				Instruction			
OP Lost		OP Won		OP Lost		OP Won		OP Lost		OP Won	
♠ Left	♠ Right	♥ Left	♥ Right	♠ Left	♠ Right	♥ Left	♥ Right	♠ Left	♠ Right	♥ Left	♥ Right

Fig. 6 Experimental design. a Time-course of a representative trial, starting with the opponent (OP) choosing one of two covered cards (1000–2500 ms; computer decision phase) and ending with the exposure of the selected card to the participant. In the participant decision phase (in a time window of 2500 ms), they had to decide what to communicate back to OP by either lying and thus changing the outcome (selecting the not-chosen card) or telling the truth and thus respecting the outcome (selecting the chosen card). The trial ended with a fixation point with a jittered inter-trial interval (1000–2500 ms). **b** The three main conditions and their symbols during the experiment. **c** Visual representation of the 3 × 2 × 2 experimental design: experimental condition (No rep, Rep, Ins)×the possible outcome [OP lost (Fav) or OP won (UnFav)]×position of the selected card by OP (left or right).

different lengths, namely, either 8, 10, or 14 trials, to avoid predictability of the number of consecutive trials. Participants received no feedback about whether they were caught lying during the functional scans. Moreover, to avoid outcome predictability, the four types of possible outcomes (Fig. 6c) were randomized across the whole experiment, and not within each mini-block. The participant could make four types of decisions depending on the outcome: OP won, i.e., unfavorable outcome (UnFav), eliciting either a self-gain lie (SG-Lie) or an other-gain truth (OG-Truth), and OP lost, i.e., favorable outcome (Fav), eliciting either an other-gain lie (OG-Lie) or a self-gain truth (SG-Truth). Instruction trials were used as control conditions in fMRI analyses.

Behavioral data analysis. To test whether Reputation (No Rep vs. Rep) and Outcome (Fav vs. UnFav) affected lying percentage, a 2 × 2 binomial generalized mixed linear model was used, with the decision as the dependent variable (lie = 1 and truth = 0) and subject as a random factor. To analyze whether response times differed between Reputation (No Rep vs. Rep), Outcome (Fav vs. UnFav) and Decisions (lie vs. truth), a 2 × 2 × 2 general linear model was used with response time as the dependent variable and subject as a random factor. Random intercepts and slopes were included in the model. Type III Wald Anova function from the R package was used to determine the statistical significance of the fixed effects for both models. Least square means (from the *lsmeans* package) and Tukey corrections were used for post-hoc comparisons of the interaction effects. Covariates of gender and age were included separately in the analysis but were ultimately excluded from the model due to their lack of significant impact.

Finally, questionnaire scores were correlated with lying percentage scores for each spontaneous condition (No Rep and Rep) using Spearman's correlation test because of the ordinal nature of the questionnaire scores. Only results passing a significance level of $p < 0.05$ are illustrated and included in the "Results" section.

Functional magnetic resonance imaging. All images were acquired with a Siemens Allegra fMRI scanner (Siemens Medical Systems, Erlangen, Germany) operating at 3 T. A quadrature volume head coil was used for radio-frequency transmission and reception. Head movements were minimized by mild restraint and cushioning. For each subject, functional MR images were acquired using echo-planar imaging (EPI) [slices = 32, TR = 2.08 s, TE = 30 ms, in-plane resolution = 3 × 3 mm^2, slice thickness = 2.5 mm, flip angle = 70°), covering the entire cortex. Structural MR images were obtained using a T1-weighted 3D magnetization prepared rapid gradient echo (MPRAGE) imaging sequence [slices = 176, TR = 2 s, TE = 4.38 ms, in-plane resolution = 0.5 × 0.5 mm^2, slice thickness = 1 mm, flip angle = 8°]. For each participant, we acquired 1284 fMRI volumes, 321 for each of the four functional runs. The first four volumes of each run were used for stabilizing longitudinal magnetization and then discarded from further analysis.

fMRI data analysis

Pre-processing. The fMRI data were pre-processed and analyzed with the Statistical Parametric Mapping package SPM12 (www.fil.ion.ucl.ac.uk) implemented in MATLAB R2019b (The MathWorks, Natick, MA). Using the ARTrepair toolbox (https://www.nitrc.org/projects/art_repair/), all images were previewed (art_movie) for detection of excessive motion artifacts, and bad slices were detected and repaired (art_slice) by interpolating adjacent slices. Functional images were then slice-time corrected to compensate for slice acquisition delays between the first and last slice by using the middle slice as a reference. Subsequently, images were realigned and unwarped to correct for head movement. Images were registered to the first volume using a 2nd-degree B-spline interpolation, whereas during the unwarp re-slicing a 4th-degree B-spline interpolation was used. Individual structural T1-weighted images were segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) using SPM tissue probability maps. Structural images were bias-corrected with a light regularization and a 60 mm cut-off, while forward deformation fields were created. The segmented structural bias-corrected image was then co-registered with the slice-timed, realigned, and unwarped functional images, using the mean unwarped image as the source image. The obtained forward deformation fields during segmentation were used during normalization to bring co-registered functional images of the 4 sessions into the MNI space (2 mm isotropic voxel) using the 4th-degree B-Spline interpolation. Finally, images were smoothed with a Gaussian kernel of 8 mm FWHM to ameliorate differences in inter-subject localization.

The pre-processed images were then examined again using the Artifact Detection Tool (2015) software package (https://www.nitrc.org/projects/artifact_detect), for detecting those scans in which—notwithstanding the above "repair" procedure—the excessive motion remained. Outlier scans were identified in the temporal differences series by assessing between-scan differences (*Z*-threshold: 3.0 mm, scan-to-scan movement threshold: 1 mm; rotation threshold: 0.02 radians). The outlier scans (3.1% overall) were omitted from the analysis by including a single regressor for each outlier in the design matrix.

fMRI general linear model analyses. The main aim of this study was to investigate how one's decision to deceive modulates neural activity under the risk of losing one's reputation. For this purpose, two separate general linear models (GLM1 and GLM2) were generated. GLM1 was performed to look at the overall neural difference between spontaneous lying and truth-telling, with an emphasis on how including a reputation risk factor modulates this difference, irrespective of the nature of the decision (self-gain or other-gain). GLM2 provides a more detailed picture of the regions involved in self-gain vs. other-gain decisions.

For each GLM, the statistical inference was based on a random-effects approach comprising two steps: first-level multiple regression models estimating contrasts of interest for each participant, followed by the second-level analyses for statistical inference at the group level, using a flexible factorial model. Non-sphericity correction[72] was applied to account for possible differences in error variance across conditions, arising—for example—because of the different number of trials in the conditions of interest and/or any non-independent error terms for the repeated measures (see second-level analyses).

Similar to previous studies on spontaneous deception[5,18,38,73,74], the naturalistic nature of the task caused an imbalance in the number of trials for both the reputation and no reputation conditions; e.g., participants who always chose other-gain truths instead of self-gain lies or vice versa. This meant that some participants had to be excluded due to an insufficient number of trials. In our first GLM, both types of lies (other-gain and self-gain) and both types of truths (other-gain and self-gain) were collapsed to contrast overall lying with truth-telling, meaning that, for example, insufficient trials for other-gain lies would be compensated with sufficient trials for self-gain lies. However, in our second GLM, all specific decisions were regressed separately meaning more imbalance and therefore, higher rates of participant exclusion.

For GLM1, data from six participants were excluded from the analyses due to insufficient trials, based on a criterion of <10% for lying or truth-telling decisions per condition (No Rep or Rep) per outcome (Fav or UnFav), i.e., 64 trials per condition meaning <6.4 trials (see Supplementary Material Table S1 for lie/truth count per condition for each participant). This left twenty-two participants (10 females, range = 20–32 years, mean age = 25.69 ± 4.12,) in GLM1. Six regressors of interest were modeled at the first level, corresponding to the following conditions: (1) spontaneous lies with no reputation risk (NoRep_Lie); (2) spontaneous truths with no reputation risk (NoRep_Truth); (3) spontaneous lies with reputation risk (Rep_Lie); (4) spontaneous truths with reputation risk (Rep_Truth); (5) instructed lies (Ins_Lie); (6) instructed truths (Ins_Truth) (see Supplementary Material Fig. S1 for a visual representation). Additionally, six sets of motion parameters derived from the realignment stage and outlier regressors were included as covariates of no interest. The events were modeled as mini blocks, time-locked to the onset of the decision phase, with a duration equal to the time window of the decision phase (i.e., 2.5 s). All regressors were convolved with the canonical haemodynamic response function (HRF), and a temporal high-pass filter with a cut-off at 128 s was applied to reduce low-frequency noise. For each participant, linear contrasts were used to average the parameter estimates associated with each of the six conditions of interest, across the four fMRI runs. For the group-level analysis, we carried out a within-subject ANOVA with factors: Condition (no rep, rep, ins) and Decision (lie, truth).

For GLM2, data from 13 participants were excluded based on the 10% criterion per decision (Self-gain Lie, Other-gain Truth, Self-Gain Truth) within each condition (No Rep or Rep) and Outcome (Fav or Unfav) (see Supplementary Material Table S2). Insufficient trials of other-gain lies (OG-Lie) were not considered since very few participants chose to lie for other-gain reasons (5.7% of NoRep trials and 5% of Rep trials across participants). For this reason, the other-gain lies were modeled at the first level but not included in the second-level model. GLM2 included a total of fifteen participants (7 females, range = 20–32 years, mean age = 25.15 ± 4.02). Twelve regressors were modeled at first level: (1) self-gain lies with no reputation risk (NoRep_SG-Lie); (2) other-gain truths with no reputation risk (NoRep_OG-Truth); (3) other-gain lie with no reputation risk (NoRep_OG-Lie); (4) self-gain truth with no reputation risk (NoRep_SG-Truth); (5) self-gain lies with reputation risk (Rep_SG-Lie); (6) other-gain truths with reputation risk (Rep_OG-Truth); (7) other-gain lie with reputation risk (Rep_OG-Lie); (8) self-gain truth with reputation risk (Rep_SG-Truth); (9) instructed self-gain lies (Ins_SG-Lie); (10) instructed other-gain truths (Ins_OG-Truth); (11) instructed other-gain lie (Ins_OG-Lie); and (12) instructed self-gain truth (Ins_OG-Truth) (see Supplementary Material Fig. S1 for a visual representation).

As in GLM1, this model included motion parameters and outlier regressors as covariates of no interest; the events were modeled as mini-blocks time-locked to the onset of the decision phase (duration = 2.5 s); all regressors were convolved with the HRF, with a cut-off filter at 128 s, and linear contrasts were used to average the parameter estimates associated with each of the twelve conditions of interest, across the four fMRI-runs. However, we did not include in the group-level analysis, the "other-gain lies" regressors for all three conditions (NoRep, Rep, and Ins) due to insufficient trial count across participants that hampers good parameter estimates. We, therefore, conducted another within-subjects ANOVA including Condition (NoRep, Rep, Ins)×Decision (SG-Lie, OG-Truth, SG-Truth).

Our contrasts of interest for GLM1 were the main effect of spontaneous lying vs. truth-telling (NoRep_Lie + Rep_Lie > NoRep_Truth + Rep_Truth) and the interaction effect of reputation and no reputation with lies and truths (Rep_Lie + NoRep_Truth > NoRep_Lie + Rep_Truth). For GLM2, we were specifically interested in clarifying the correlates of SG-Lies vs. OG-Truth for both Rep and NoRep conditions in unfavorable outcomes.

For both GLM1 and GLM2, in line with the main aim of the study, we constrained the search volume (using the small volume correction SPM function) within the brain areas responding to spontaneous decisions, i.e., all "spontaneous vs. instructed" trials, using a thresholded contrast image of $p_{\text{uncorrected}} < 0.005$ (see Fig. S5). Data are presented including all significant activations at peak-level using family-wise corrected p-values (significance at $P_{\text{FWE-corrected}} < 0.05$). Whole-brain analysis results for both GLM1 and GLM2 are included in the supplementary materials (Table S5, S6, S7, S8).

For both models, covariates of gender and age were included as covariate regressors but were ultimately excluded from the model due to their lack of significant impact.

Questionnaires. After the task and outside the fMRI scanner, the participants were qualitatively debriefed about their experience. A questionnaire was given right after the scanning session (see Supplementary Material Q1). Two participants declared they did not believe that the OP was a real player and thus were excluded from the analysis. After this, the participants were administered: the Balanced Inventory of Desirable Responding (BIDR)[75], which consists of two 20-item subscales, ranging from 20–140 and measures self-deception and impression management, both related to social desirability; the Machiavellianism Scale (MACH-IV)[76], which is a 20-item scale where scores can range from 40 to 160 and measures the ability to use deception and manipulation to acquire power during everyday life interactions; and the Civic Moral Disengagement (CMD)[77], a 40-item questionnaire, scoring from 40 to 200, and measures an individual's tendency to make use of self-dismissal when violating civic duties and obligations in order to soften the moral consequences of their behavior[78]. Questionnaires were chosen based on the study of Panasiti et al. (2011)[30].

Correlation analyses. A Spearman's correlation analysis was conducted to evaluate whether bold activity in the brain regions highlighted by GLM1 and GLM2 were correlated with questionnaire measures (i.e., MACH-IV, BIDR, and CMD). Questionnaire scores were correlated with the bold signal extracted eigenvariate values for each condition. For this, we used SPM12 and created an 8 mm sphere (matching the Gaussian kernel) centered around all significant peaks that survived family-wise error correction during the GLM2 analyses. In the main text, we included correlations of regions of interest that are specifically linked to SG-Lies (GLM2) when reputation was at risk. All other correlations are available in the supplementary materials.

Task-based connectivity analyses. Finally, we explored task-modulated functional connectivity to investigate potential connectivity differences in regions uncovered in GLM1 and GLM2 within the restricted search volume mentioned above. To this end, we used a generalized psycho-physical interaction (gPPI) approach implemented through the CONN toolbox (www.nitrc.org/projects/conn). For each participant, we constructed two different PPI–GLMs (one based on each GLM). Regions of interest (ROIs), defined based on our GLM analyses (see Table S3), were established using 8-mm spheres (matching the Gaussian kernel and built by the SPM toolbox MarsBaR; Brett et al., 2002) centered on the peak voxels from significant clusters of the contrasts (see Table 4.1a & b and Table S4.1a). For each ROI, bivariate regression matrices were calculated, yielding standardized regression coefficients that estimated the functional connectivity at the group level. Only results with FDR-corrected p-values < .05 are discussed.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The fMRI datasets of both models (source data: betas and contrast nifty files and SPM.mat files) (https://doi.org/10.17605/OSF.IO/ZVKNC) and behavioral data (source data: excel behavioral data matrix) (https://doi.org/10.17605/OSF.IO/DX8JU) that support the findings of this study are available as a repository on the Open Science Framework: https://osf.io/v2jpn/). The repository also contains the raw MATLAB experimental output files and the stimuli images (https://doi.org/10.17605/OSF.IO/YK3NX). Raw fMRI data, preprocessed nifti files and first-level SPM data will be made available upon request.

Code availability

R./RStudio was used for data analysis. The R. code and the corresponding excel data matrix can be found here: https://doi.org/10.17605/OSF.IO/DX8JU. For data collection, MATLAB R2019b was used. MATLAB code for running the experiment and extracting trial-by-trial data can be found here: https://doi.org/10.17605/OSF.IO/YK3NX. MATLAB code for extracting onsets for fMRI data analysis can be found here: https://doi.org/10.17605/OSF.IO/ZVKNC.

Received: 28 November 2022; Accepted: 6 April 2023;
Published online: 29 April 2023

References

1. Ganis, G. Neural correlates of different types of deception: an fMRI investigation. *Cereb. Cortex* **13**, 830–836 (2003).
2. Jiang, W. et al. Decoding the processing of lying using functional connectivity MRI. *Behav. Brain Funct.* **11**, 1 (2015).
3. Pardo, M. S. Lying, deception, and fMRI: a critical update. In *Neurolaw and Responsibility for Action* (ed. Donnelly-Lazarov, B.) 143–159 (Cambridge University Press, 2018).
4. Volz, K. G., Vogeley, K., Tittgemeyer, M., von Cramon, D. Y. & Sutter, M. The neural basis of deception in strategic interactions. *Front. Behav. Neurosci.* **9**, 27 (2015).
5. Yin, L., Reuter, M. & Weber, B. Let the man choose what to do: Neural correlates of spontaneous lying and truth-telling. *Brain Cogn.* **102**, 13–25 (2016).
6. Yin, L. & Weber, B. I lie, why don't you: neural mechanisms of individual differences in self-serving lying. *Hum. Brain Mapp.* **40**, 1101–1113 (2019).
7. Yu, J., Tao, Q., Zhang, R., Chan, C. C. H. & Lee, T. M. C. Can fMRI discriminate between deception and false memory? A meta-analytic comparison between deception and false memory studies. *Neurosci. Biobehav. Rev.* **104**, 43–55 (2019).
8. Zheltyakova, M., Kireev, M., Korotkov, A. & Medvedev, S. Neural mechanisms of deception in a social context: an fMRI replication study. *Sci. Rep.* **10**, 1–12 (2020).
9. Lee, T. M. C. et al. Lie detection by functional magnetic resonance imaging. *Hum. Brain Mapp.* **15**, 157–164 (2002).
10. Abe, N. et al. The neural basis of dishonest decisions that serve to harm or help the target. *Brain Cogn.* **90**, 41–49 (2014).
11. Ding, X. P., Wu, S. J., Liu, J., Fu, G. & Lee, K. Functional neural networks of honesty and dishonesty in children: evidence from graph theory analysis. *Sci. Rep.* **7**, 12085 (2017).
12. Greene, J. D. & Paxton, J. M. Patterns of neural activity associated with honest and dishonest moral decisions. *Proc. Natl Acad. Sci. USA* **106**, 12506–12511 (2009).
13. Pornpattananangkul, N., Zhen, S. & Yu, R. Common and distinct neural correlates of self-serving and prosocial dishonesty. *Hum. Brain Mapp.* **39**, 3086–3103 (2018).
14. Sun, D., Chan, C. C. H., Hu, Y., Wang, Z. & Lee, T. M. C. Neural correlates of outcome processing post dishonest choice: an fMRI and ERP study. *Neuropsychologia* **68**, 148–157 (2015).
15. Speer, S., Smidts, A. & Boksem, M. A. S. Individual differences in (dis)honesty are represented in the brain's functional connectivity at rest. *NeuroImage* **246**, 118761 (2022).
16. Speer, S., Smidts, A. & Boksem, M. A. S. Cognitive control promotes either honesty or dishonesty, depending on one's moral default. *J. Neurosci.* **41**, 8815–8825 (2021).
17. Ofen, N., Whitfield-Gabrieli, S., Chai, X. J., Schwarzlose, R. F. & Gabrieli, J. D. E. Neural correlates of deception: lying about past events and personal beliefs. *Soc. Cogn. Affect. Neurosci.* **12**, 116–127 (2017).
18. Lisofsky, N., Kazzer, P., Heekeren, H. R. & Prehn, K. Investigating socio-cognitive processes in deception: a quantitative meta-analysis of neuroimaging studies. *Neuropsychologia* **61**, 113–122 (2014).
19. Mameli, F. et al. Honesty. In *Neuroimaging personality, social cognition, and character*. (eds Absher, J. R. & Cloutier, J.) 305–322 (Academic Press, 2016).
20. Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E. & McDermott, K. B. The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cereb. Cortex* **19**, 1557–1566 (2009).
21. Speer, S., Smidts, A. & Boksem, M. A. S. Cognitive control and dishonesty. *Trends Cogn. Sci.* **26**, 796–808 (2022).
22. Ito, A. et al. The role of the dorsolateral prefrontal cortex in deception when remembering neutral and emotional events. *Neurosci. Res.* **69**, 121–128 (2011).
23. Abe, N., Suzuki, M., Mori, E., Itoh, M. & Fujii, T. Deceiving others: distinct neural responses of the prefrontal cortex and amygdala in simple fabrication and deception with social interactions. *J. Cogn. Neurosci.* **19**, 287–295 (2007).
24. Sip, K. E. et al. When Pinocchio's nose does not grow: belief regarding lie-detectability modulates production of deception. *Front. Hum. Neurosci.* **7**, 16 (2013).
25. Delgado-Herrera, M., Reyes-Aguilar, A. & Giordano, M. What deception tasks used in the lab really do: systematic review and meta-analysis of ecological validity of fMRI deception tasks. *Neuroscience* **468**, 88–109 (2021).
26. Sip, K. E., Roepstorff, A., McGregor, W. & Frith, C. D. Detecting deception: the scope and limits. *Trends Cogn. Sci.* **12**, 48–53 (2008).
27. Garrett, N., Lazzaro, S. C., Ariely, D. & Sharot, T. The brain adapts to dishonesty. *Nat. Neurosci.* **19**, 1727–1732 (2016).
28. Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K. & Fehr, E. The neural circuitry of a broken promise. *Neuron* **64**, 756–770 (2009).
29. Panasiti, M. S. et al. Thermal signatures of voluntary deception in ecological conditions. *Sci. Rep.* **6**, 35174 (2016).
30. Panasiti, M. S., Pavone, E. F., Merla, A. & Aglioti, S. M. Situational and dispositional determinants of intentional deceiving. *PLoS ONE* **6**, e19465 (2011).
31. Panasiti, M. S. et al. The motor cost of telling lies: electrocortical signatures and personality foundations of spontaneous deception. *Soc. Neurosci.* 1–17 https://doi.org/10.1080/17470919.2014.934394 (2014).

32. Azevedo, R. T., Panasiti, M. S., Maglio, R. & Aglioti, S. M. Perceived warmth and competence of others shape voluntary deceptive behaviour in a morally relevant setting. *Br. J. Psychol.* **109**, 25–44 (2018).
33. Vabba, A., Porciello, G., Panasiti, M. S. & Aglioti, S. M. Interoceptive influences on the production of self-serving lies in reputation risk conditions. *Int. J. Psychophysiol.* **177**, 34–42 (2022).
34. Schepisi, M., Porciello, G., Aglioti, S. M. & Panasiti, M. S. Oculomotor behavior tracks the effect of ideological priming on deception. *Sci. Rep.* **10**, 9555 (2020).
35. Vonasch, A. J., Reynolds, T., Winegard, B. M. & Baumeister, R. F. Death before dishonor: incurring costs to protect moral reputation. *Soc. Psychol. Personal. Sci.* **9**, 604–613 (2018).
36. Azevedo, R. T. et al. Their pain is not our pain: brain and autonomic correlates of empathic resonance with the pain of same and different race individuals. *Hum. Brain Mapp.* **3181**, 3168–3181 (2013).
37. Sisco, M. R. & Weber, E. U. Examining charitable giving in real-world online donations. *Nat. Commun.* **10**, 3968 (2019).
38. Sip, K. E. et al. What if I get busted? Deception, choice, and decision-making in social interaction. *Front. Neurosci.* **6**, 1–10 (2012).
39. Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G. & Crombez, G. Lying takes time: A meta-analysis on reaction time measures of deception. *Psychol. Bull.* **143**, 428–453 (2017).
40. Yin, L., Hu, Y., Dynowski, D., Li, J. & Weber, B. The good lies: altruistic goals modulate processing of deception in the anterior insula. *Hum. Brain Mapp.* https://doi.org/10.1002/hbm.23623 (2017).
41. Meier, S. K., Ray, K. L., Mastan, J. C., Salvage, S. R. & Robin, D. A. Meta-analytic connectivity modelling of deception-related brain regions. *PLoS ONE* **16**, e0248909 (2021).
42. Lavin, C. et al. The anterior cingulate cortex: an integrative hub for human socially-driven interactions. *Front. Neurosci.* **7**, 64 (2013).
43. Abe, N., Greene, J. D. & Kiehl, K. A. Reduced engagement of the anterior cingulate cortex in the dishonest decision-making of incarcerated psychopaths. *Soc. Cogn. Affect. Neurosci.* **13**, 797–807 (2018).
44. Menon, V., Adleman, N. E., White, C. D., Glover, G. H. & Reiss, A. L. Error-related brain activation during a Go/NoGo response inhibition task. *Hum. Brain Mapp.* **12**, 131–143 (2001).
45. Posner, M. I. & Petersen, S. E. The attention system of the human brain. *Annu. Rev. Neurosci.* **13**, 25–42 (1990).
46. Smith, R. et al. Maintaining the feelings of others in working memory is associated with activation of the left anterior insula and left frontal-parietal control network. *Soc. Cogn. Affect. Neurosci.* **12**, 848–860 (2017).
47. Stemmer, B., Vihla, M. & Salmelin, R. Activation of the human sensorimotor cortex during error-related processing: a magnetoencephalography study. *Neurosci. Lett.* **362**, 44–47 (2004).
48. Craig, A. Interoception: the sense of the physiological condition of the body. *Curr. Opin. Neurobiol.* **13**, 500–505 (2003).
49. Ham, T., Leff, A., de Boissezon, X., Joffe, A. & Sharp, D. J. Cognitive control and the salience network: an investigation of error processing and effective connectivity. *J. Neurosci.* **33**, 7091–7098 (2013).
50. Menon, V. & Uddin, L. Q. Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* **214**, 655–667 (2010).
51. Korucuoglu, O. et al. Test-retest reliability of fMRI-measured brain activity during decision making under risk. *NeuroImage* **214**, 116759 (2020).
52. Loued-Khenissi, L., Pfeuffer, A., Einhäuser, W. & Preuschoff, K. Anterior insula reflects surprise in value-based decision-making and perception. *NeuroImage* **210**, 116549 (2020).
53. Sai, L., Wu, H., Hu, X. & Fu, G. Telling a truth to deceive: examining executive control and reward-related processes underlying interpersonal deception. *Brain Cogn.* **125**, 149–156 (2018).
54. Craig, A. How do you feel—now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* **10**, 59–70 (2009).
55. Eisenberger, N. I. Does rejection hurt? An fMRI study of social exclusion. *Science* **302**, 290–292 (2003).
56. Sanfey, A. G. The neural basis of economic decision-making in the ultimatum game. *Science* **300**, 1755–1758 (2003).
57. Zaki, J. & Mitchell, J. P. Equitable decision making is associated with neural markers of intrinsic value. *Proc. Natl Acad. Sci. USA* **108**, 19761–19766 (2011).
58. Panasiti, M. S. et al. Cognitive load and emotional processing in psoriasis: a thermal imaging study. *Exp. Brain Res.* **237**, 211–222 (2019).
59. Seara-Cardoso, A., Sebastian, C. L., Viding, E. & Roiser, J. P. Affective resonance in response to others' emotional faces varies with affective ratings and psychopathic traits in amygdala and anterior insula. *Soc. Neurosci.* **11**, 140–152 (2016).
60. Mohr, P. N. C., Biele, G. & Heekeren, H. R. Neural processing of Risk. *J. Neurosci.* **30**, 6613–6619 (2010).
61. Wu, C. C., Sacchet, M. D. & Knutson, B. Toward an affective neuroscience account of financial risk taking. *Front. Neurosci.* **6**, 159 (2012).
62. Rudorf, S., Preuschoff, K. & Weber, B. Neural correlates of anticipation risk reflect risk preferences. *J. Neurosci.* **32**, 16683–16692 (2012).
63. Bellucci, G., Feng, C., Camilleri, J., Eickhoff, S. B. & Krueger, F. The role of the anterior insula in social norm compliance and enforcement: evidence from coordinate-based and functional connectivity meta-analyses. *Neurosci. Biobehav. Rev.* **92**, 378–389 (2018).
64. Paulus, M. P. & Stein, M. B. An insular view of anxiety. *Biol. Psychiatry* **60**, 383–387 (2006).
65. Phan, K. L., Wager, T., Taylor, S. F. & Liberzon, I. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage* **16**, 331–348 (2002).
66. Shin, L. M. et al. Activation of anterior paralimbic structures during guilt-related script-driven imagery. *Biol. Psychiatry* **48**, 43–50 (2000).
67. Sip, K. E. et al. What if I get busted? Deception, choice, and decision-making in social interaction. *Front. Neurosci.* **6**, 58 (2012).
68. Venkatraman, V., Rosati, A. G., Taren, A. A. & Huettel, S. A. Resolving Response, Decision, and Strategic Control: Evidence for a Functional Topography in Dorsomedial Prefrontal Cortex. *J. Neurosci.* **29**, 13158–13164 (2009).
69. Venkatraman, V. & Huettel, S. A. Strategic control in decision-making under uncertainty: Strategic control in decision making. *Eur. J. Neurosci.* **35**, 1075–1082 (2012).
70. Cui, F. et al. Altruistic and self-serving goals modulate behavioral and neural responses in deception. *Soc. Cogn. Affect. Neurosci.* **13**, 63–71 (2018).
71. Zheltyakova, M., Kireev, M., Korotkov, A. & Medvedev, S. Neural mechanisms of deception in a social context: an fMRI replication study. *Sci. Rep.* **10**, 10713 (2020).
72. Friston, K. J. et al. Classical and Bayesian inference in neuroimaging: applications. *NeuroImage* **16**, 484–512 (2002).
73. Lin, X. A., Wang, C., Zhou, J., Sai, L. & Fu, G. Neural correlates of spontaneous deception in a non-competitive interpersonal scenario: a functional near-infrared spectroscopy (fNIRS) study. *Brain Cogn.* **150**, 105704 (2021).
74. Sip, K. E. et al. The production and detection of deception in an interactive game. *Neuropsychologia* **48**, 3619–3626 (2010).
75. Paulhus, D. L. & Reid, D. B. Enhancement and denial in socially desirable responding. *J. Personal. Soc. Psychol.* **60**, 307–317 (1991).
76. Christie, R. & Geis, F. L. *Studies in Machiavellianism* (Academic Press, New York, 1970).
77. Bandura, A., Barbaranelli, C., Caprara, G. V. & Pastorelli, C. Mechanisms of moral disengagement in the exercise of moral agency. *J. Personal. Soc. Psychol.* **71**, 364 (1996).
78. Caprara, G. V., Fida, R., Vecchione, M., Tramontano, C. & Barbaranelli, C. Assessing civic moral disengagement: dimensionality and construct validity. *Personal. Individ. Differ.* **47**, 504–509 (2009).

Acknowledgements

This work was supported by a European Research Council (ERC) Advanced Grant 2017 awarded to SMA (grant number ERC-2017-AdG—eHONESTY—789058). The authors would like to thank Marco Tullio Liuzza and Emiliano Macaluso for their involvement in the preliminary stage of the study.

Author contributions

Conceptualization, S.M.A., R.T.A., M.S.P.; Methodology and analysis, L.D., V.S., R.T.A. Writing—original draft, L.D., V.S., S.M.A.; Writing—review & editing, all the authors; Funding acquisition, S.M.A.; Resources, S.M.A.; Supervision S.M.A. and V.S.

Competing interests

The authors declare no competing interests.

Inclusion and diversity

We worked to ensure sex balance in the recruitment of human subjects and to prepare the study questionnaires in an inclusive way.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-023-04827-w.

Correspondence and requests for materials should be addressed to Lennie Dupont or Salvatore Maria Aglioti.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Karli Montague-Cardoso.

Reprints and permission information is available at http://www.nature.com/reprints