CONTRIBUTION OF SELF- AND OTHER-REGARDING MOTIVES TO (DIS)HONESTY

*Anastasia Shuster[1,2] and Dino J Levy[1,2]

[1]*Sagol School of Neuroscience, Tel Aviv University*

[2]*Coller School of Management, Tel Aviv University*

Acknowledgements. We thank Liz Izakson for her help in data collection, and Dr. Xiaosi Gu for her thoughtful comments on the manuscript. The project was funded by a BSF-NSF grant #0612015334 and ISF grant #0612015023 awarded to DJL.

Author contributions. DJL oversaw the project. AS and DJL designed the experiment. AS collected and analyzed the data. AS and DJL wrote the manuscript.

Competing interests. The authors declare no competing interests.

Materials & Correspondence. Correspondence and material requests should be addressed to Anastasia Shuster, Anastasia.Shuster@mssm.edu.

# Pages: 39

# Words (main): 6,798

# Figures: 5

# Tables: 2

# Supplementary material: 1 table, 3 figures

Abstract

Why would people tell the truth when there is an obvious gain in lying and no risk of being caught? Previous work suggests the involvement of two motives, self-interest and regard for others. However, it remains unknown if these are related or independently contribute to dishonesty. Using a modified Message Game task, in which a Sender sends a dishonest (yet profitable) or honest (less profitable) message to a Receiver, we found that these two motives contributed to dishonesty independently. Furthermore, distinct brain networks represented the two motives: the LPFC and ACC tracked potential value to self, whereas the rTPJ, vmPFC, and ventral striatum tracked potential losses to other. Individual differences in motives modulated these neural responses. Finally, vmPFC activity represented motive-modulated integration of values to self and other. Taken together, our results suggest that (dis)honest decisions incorporate at least two separate cognitive and neural processes – valuation of potential profits to self and valuation of potential harm to others.

Introduction

"Children and fools tell the truth" — proverb

"Thou shalt not bear false witness against thy neighbor" — Exodus 20:16, The Bible

Honesty is a social norm, and yet people lie quite often (Vrij, 2004). Previous research suggests that the decision to lie may depend on (1) the size of the profit gained from lying (*self-interest*), and (2) the degree of harm that the lie would cause (*regard for others*) (Gneezy, 2005). Other self-related motives, such as the chance of being caught (Mazar et al., 2008), the wish to maintain a positive self-image (Mazar et al., 2008; Jacobsen et al., 2018), and an aversion to lying (Gneezy et al., 2013) would also decrease dishonesty. However, less research has looked into how *outcomes* of dishonest behavior affect it (Gneezy, 2005). Moreover, the neural computation underlying the arbitration between self- and other-regarding motives remains elusive.

We used a modified version of the Message Game, in which a *Sender* sends either a truthful or a deceiving message to a *Receiver* regarding which of two options to choose. The task conflicts monetary gain to the Sender with honesty, to invoke internally motivated lying. As potential profits to the Sender go up, Senders tend to send the deceptive message more often. Conversely, as potential losses to the Receivers rise, Senders lie less (Gneezy, 2005). By systematically varying the payoffs to both players, we can estimate the role of self- and other-regarding motives in dishonest choices and track their neural correlates. In the original Message Game, Sender's message does not always predict the Receiver's choice (Gneezy, 2005; Gneezy et al., 2013). Because there are only two options, a particularly untrusting Receiver could choose the opposite option than the one recommended to her by the Sender. This opens the door to strategic choices, in which a Sender might choose to tell the truth while having the intention to deceive (Volz et al., 2015). Therefore, in the current study, we modified the task to include two additional options per trial, in which both players (Sender and Receiver) stood to gain $0. Thus, deviating from the Sender's message would result in a 66% chance of not winning any money at all. This important modification of the task ensures that Senders' choices have true consequences for their partners.

Neuroimaging studies consistently implicate several brain regions in the prefrontal cortex (PFC) in generating dishonest behavior (Spence et al., 2001; Abe, 2009, 2011; Christ et al., 2009; Greene and Paxton, 2009; Lisofsky et al., 2014). For example, deceptive responses—but not erroneous ones—selectively activate the middle frontal gyrus (Abe, 2009; Liang et al., 2012). Purposefully withholding information from the experimenter engages the anterior cingulate cortex (ACC), dorsolateral prefrontal cortex (dlPFC), inferior and superior frontal gyrus (Kozel et al., 2004, 2005; Langleben et al., 2005; Bhatt et al., 2009). Structural examination of pathological liars' brains revealed up to 36% more white matter in various parts of the prefrontal cortex compared to controls (Yang et al., 2005, 2007). Not surprisingly, the involvement of the prefrontal cortex has raised the possibility that lying requires excursion of cognitive control (Spence, 2004), supported by substantial overlap between areas of the brain implicated in deception and areas related to executive control (Christ et al., 2009).

Much of the previous neuroimaging studies explicitly instructed participants when to lie, and therefore lack the external validity needed to probe *internally motivated dishonesty* (Lisofsky et al., 2014; Yin et al., 2016). Moreover, they involved withholding information from the experimenter or deceiving her, but dishonest behavior did not come at the expense of another person. As such, they could not address other-regarding motives as drivers of dishonesty or search for their neural correlates.

Interestingly, a meta-analysis comparing interactive (i.e., involving another participant) with non-interactive deception studies revealed that the neural activity associated with interactive deception is different from that of non-interactive deception, and marked by greater activations in the dorsal ACC, TPJ and temporal poles (Lisofsky et al., 2014). Therefore, a valid and interactive task is required to study motives for (dis)honest behavior. Recently, a group of researchers put forward a signaling framework to study deception, drawing from game theory (Jenkins et al., 2016). Using the Message Game, which is a type of a signaling task, researchers have demonstrated that lesions to the dlPFC reduce honesty concerns in favor of self-interest (Zhu et al., 2014). Thus, the prefrontal cortex appears to have a direct causal role in volitional and interactive dishonest behavior. However, studies of neural correlates of deception did not look into the social outcomes of deception, and how other-regarding preferences may be involved.

Evidence for the integration of other-regarding motives with self-interest come from studies of prosocial behavior. Even in the absence of explicit extrinsic pressure, humans are willing to forego monetary gain to cooperate (Rilling et al., 2002), donate and share resources (Andreoni and Miller, 2002; Moll et al., 2006; Hare et al., 2010) and act fairly (Zaki and Mitchell, 2011). On the neural level, acting prosocially activates areas of the neural valuation system, consisting of the ventromedial prefrontal cortex (vmPFC) and ventral striatum (VS) (Moll et al., 2006; Zaki and Mitchell, 2011; for reviews see Fehr and Camerer, 2007; Ruff and Fehr, 2014). Experiencing vicarious reward engages the vmPFC (Zaki et al., 2014; Morelli et al., 2015), and choosing on behalf of another activates the vmPFC in a similar fashion to when choosing for oneself (Janowski et al., 2013). Taken together, these findings indicate a common computation of value in these regions, where both social (e.g., norms) and nonsocial (e.g., monetary profit) factors are integrated into a common decision-value, which gives rise to choice. Putatively, when choices present a conflict between one's own profit and normative social principles, the valuation system interacts with areas typically involved in social cognition (e.g., the TPJ (Smith et al., 2013) or pSTS (Hare et al., 2010)) to compute the subjective value of an alternative (Ruff and Fehr, 2014). Thus, in the present study, we examine the neural correlates of the *drivers* of (dis)honest behavior. We focus on the unique contribution of self- and other-regarding motives for (dis)honest behavior, as well as the integration of the two in the valuation system.

Participants played as Senders, choosing between honest and dishonest alternatives. Dishonesty was associated with higher gain for themselves and greater losses to their partner. Our first aim was to measure how increases to own profit and other's loss affect dishonest choices. Our second aim was to explore individual differences in these motives. Third, we aimed to identify the neural activity associated with the two value parameters that drive dishonest behavior – value to self and other. Finally, our fourth aim was to examine how the neural representation of these value parameters reflects individual differences in behavior.

We predicted that dishonest behavior would engage brain areas previously identified in the literature: the ACC, dlPFC, ventrolateral prefrontal cortex (vlPFC), insula, and inferior parietal cortex (IPL) (Kozel et al., 2005; Christ et al., 2009; Lisofsky et al., 2014). We further predicted specific neural representations for the two value parameters; the first, value to self, operationalized as the Sender's potential profits from sending a deceiving message; and second, value to other – the potential losses to the Receiver inflicted by such dishonesty. We expected that dishonest profits to self would implicate the dlPFC (Zhu et al., 2014; Crockett et al., 2017), and that losses to the Receiver would activate the social cognition network, namely, the TPJ and/or the temporal pole (Carter et al., 2012; Lisofsky et al., 2014). The TPJ and the adjacent IPL have been implicated in several deception studies (Christ et al., 2009; Lisofsky et al., 2014), and in social cognition (Saxe and Kanwisher, 2003; Carter et al., 2012). This makes it a prime candidate for representing other-regarding motives. Finally, the vmPFC was of special interest, due it its role in valuation (Levy and Glimcher, 2012; Bartra et al., 2013) and its involvement in prosocial decision-making (Fehr and Camerer, 2007; Ruff and Fehr, 2014). Because the vmPFC represents both own and vicarious/social rewards (Zaki and Mitchell, 2011; Janowski et al., 2013; Morelli et al., 2015), we hypothesized that it would represent the integration of value to self and other.

Results

Behavior

On each trial in the task, the participant (*Sender*) chose to send either a truthful or a deceptive message to the *Receiver* (see Fig. 1). We started with a simple measure of overall dishonesty – on how many of the trials did participants choose to send the deceptive message. The deceptive message was sent on almost half of the trials, but with substantial variability between participants (*M*=45.45%, *SD*=17.7%; range: 17.36%-89.93%; Fig. 2a). Overall dishonesty did not differ statistically between female and male participants (females: *M*=43% *SD*=15.9% males: *M*=50.59%, *SD*=21%; *t*(27)=-1.06, *p*=0.29, two-tailed two-sample t-test). Participants took on average 2.87s to choose, and reaction times for Truth choices (*M*=2.86 s *SD*=0.45 s) did not differ from reaction times for Lie choices (*M*=2.88 s *SD*=0.55 s) (t(27)=0.4, *p*=0.69). The difference in honest and dishonest decision times, however, is significantly affected by individual differences in overall dishonesty: participants who lied more took longer to tell the truth, whereas more honest participants took longer to lie (r(26)=-0.8, *p*<0.0001; Figure 2b). This suggests that the decision process reflects individual differences in dishonesty preference, both in which alternative is chosen and in the time it takes to choose.

Motives for (dis)honesty

To uncover what drove dishonest behavior, we conducted a multiple linear regression analysis of the probability to lie as a function of the potential payoffs, separately for each participant. The regression revealed that both the potential profits for the Sender (*value to self, ΔV_{self}*) and the potential losses to the Receiver (*value to other, ΔV_{other}*) affected the behavior of most participants (Figure 2a). A large ΔV_{self} coefficient implies a more *self-interested* participant. In other words, each unit of money to the

Sender would cause a bigger increase in the probability to lie, compared to a small ΔV_{self} coefficient. Similarly, a large $\Delta V_{\mathrm{other}}$ coefficient (high *regard for others*) means that the loss to the Receiver greatly decreases the probability of the Sender to lie, compared to a small coefficient. While the average contribution of both parameters was similar in absolute terms ($\beta\Delta V_{\mathrm{self}}$: *M*=0.174 *SD*=0.06, $\beta\Delta V_{\mathrm{other}}$: *M*=-0.169 *SD*=0.057, t(27)=0.35, *p*=0.73, paired two-tailed t-test), they varied substantially between participants ($\beta\Delta V_{\mathrm{self}}$ range: 0.06-0.29; $\beta\Delta V_{\mathrm{other}}$ range: -0.28-0.004). Our finding that both coefficients are significant implies that self-interest and regard for others independently affect the probability of the Sender to lie. This notion of independence between the two motives is strengthened by a lack of correlation between the two coefficients across participants (r(26)=-0.2, *p*=0.3; Figure S2). Finally, overall dishonesty across participants only marginally correlated with the regard for others' coefficients (r(26)=-0.35, *p*=0.062), and did not correlate with the self-interest coefficients (r(26)=0.28, *p*=0.14).

Neuroimaging

Neural correlates of dishonesty

To identify neural correlates of dishonest behavior, we contrasted the neural response during trials in which participants lied (i.e., sent the deceptive message) with trials in which they told the truth, controlling for reward amount. Consistent with previous studies of deception (Christ et al., 2009; Lisofsky et al., 2014), we found several regions, including the medial PFC, left dlPFC, and bilateral insula, to be more active during lying compared to truth-telling. The opposite comparison (Truth>Lie) revealed activations in the right TPJ, right STS and cerebellum (see Table 1 and Figure S3).

Chosen value representation

On each trial, the participant made a choice between two alternatives, each holding some amount of money for her to gain. To identify voxels responding to the Sender's reward magnitude in the chosen alternative vs. the unchosen one, we parametrically modelled the Sender's expected reward based on her choice. Consistent with previous findings (Levy and Glimcher, 2012; Bartra et al., 2013), the amount of money in the chosen vs. the unchosen option for the Sender positively correlated with the BOLD signal in the valuation system – the vmPFC and bilateral ventral striatum (Figure 3a).

Value representation for self and other

Value to Self. We found that the left LPFC and IPL, among other regions, negatively tracked the amount of money a Sender can gain from lying (ΔV_{self}; Table 1, Fig. 3b). That is, a smaller potential profit from lying (so-called ill-gotten gains, (Crockett et al., 2017)) corresponds to a higher activation in the left LPFC and IPL. To examine how individual differences in self-interest affect this representation, we extracted the BOLD coefficients for ΔV_{self} from an independently defined ROI taken from a previous study (MNI coordinates x, y, z: -48, 6, 28; (Crockett et al., 2017)) and correlated them with $\beta\Delta V_{\text{self}}$ estimated from participants' behavior. We found a positive significant correlation (r(26)=0.42, *p*=0.027), where less self-interested participants (i.e., have higher honesty concerns) show more deactivations of the LPFC.

A potential explanation to this pattern of results comes from the role of the LPFC in cognitive control (Badre and Nee, 2018) – participants might experience more conflict and need for control when the potential reward from lying ($\$\text{Self}_{\text{Lie}}$-$\$\text{Self}_{\text{Truth}}$) is small. If this was the case, we would expect to find a negative relationship between the potential reward from lying and reaction times, such that smaller differences would yield longer reaction times. Because our neural model controls for the effects of reaction times on value representation, it is a less plausible explanation for the observed result. Nonetheless, we directly tested this hypothesis behaviorally, by regressing ΔV_{self} onto reaction times (while clustering the errors per participant). We find no significant relationship between reaction times and ΔV_{self} ($\beta\Delta V_{self}$=0.005, *p*=0.48). Furthermore, each participant's correlation coefficient between reaction times and ΔV_{self} is unrelated to their $\beta\Delta V_{self}$ (r(27)=-0.02, *p*=0.89). That is, the relationship between potential profits and reaction times does not predict levels of self-interest. Interestingly, it is negatively linked to overall dishonesty (r(27)=-0.74, *p*<0.001), such that honest participants take longer to choose when they can gain a large profit from lying, while dishonest participants take longer to choose when the potential profit from lying is small.

Value to Other. To identify voxels representing the value the Sender exerts toward the Receiver (i.e., value to other, ΔV_{other}), we regressed the potential monetary loss to the Receiver if the Sender chooses to act dishonestly ($\$Other_{Lie}$-$\$Other_{Truth}$). We found an inverse relationship between the activity of the rTPJ and Receiver's potential loss, such that higher potential losses to the Receiver *deactivated* the rTPJ (Table 1; Fig. 3c). This seems counter intuitive, as the TPJ is part of the neural social cognition network. We extracted neural responses from the TPJ using an independently defined ROI taken from a previous study (MNI coordinates x, y, z: -54, -58, 22; (Crockett et al., 2017)) and examined whether differences between participants in the activity of the rTPJ can be explained by individual differences in their other-regarding motive for lying. We found that participants with low regard for others showed rTPJ deactivation, whereas in those who have high regard for others, higher potential losses elicit *more* activity in the rTPJ (r(25)=0.51, *p*=0.006). That is, the degree to which social consequences affect an individual's behavior is related to how these social consequences are represented in the rTPJ.

Individual differences modulate value representation

Participants in the study exhibited substantial individual differences in the weight they place on their own ($\beta\Delta V_{self}$) and their partner's ($\beta\Delta V_{other}$) monetary profits and losses. We directly examined how the neural representations of value reflects these behavioral individual differences in participants' motives for (dis)honesty. To this end, we conducted two analyses. First, we regressed individual differences in self-interest onto a whole-brain group-level map of value to self. Second, we regressed individual differences in regard for others onto a map of value to other. This yielded two correlation maps. For self-interest, we found that the ACC, among other regions, represents ΔV_{self}, positively modulated by $\beta\Delta V_{self}$ (Table 2). Such that, the more a person cares about their own monetary wellbeing, the more sensitive her ACC is to potential monetary profits. As for regard for others, the rTPJ, vmPFC and ventral striatum represent ΔV_{other}, positively modulated by $\beta\Delta V_{other}$. That is, participants who place greater weight on others, also have a greater neural response to the other's potential losses in these areas, compared to participants who do not care as much about others' monetary wellbeing. This suggests that each of the two behavioral motives for (dis)honesty can be traced to distinct neural processes affecting the neural representation of value.

Balanced self-other representation

To understand how the vmPFC integrates the competing values and motives, we directly analyzed the valuation system for differences between representations of value to self and value to other, and how self- and other-regarding motives affect such representations. We examined whether the vmPFC and/or ventral striatum represent the relative value of other and self, and whether it is modulated by the *relative contribution* of each motive to behavior. To do so, we first computed for each participant an *other-self differential score* ($\beta \Delta V_{other} – \square \Delta V_{self}$), indicating how much more one motive drives the participant's behavior compared to the other. High positive scores indicate a higher contribution of other-regarding motives, whereas a high negative score suggests higher contribution of self-regarding motives, and participants with scores close to zero place similar weights on self and other. Then, we directly compared the valuation system's neural representation of value to other with that of value to self ($\Delta V_{other} > \Delta V_{self}$). This contrast yielded an empty map. However, when we regressed onto this map each participant's other-self differential score, we identified a cluster of voxels located in the vmPFC (MNI coordinate x, y, z: 9, 41, -8; Figure 5). We found that the vmPFC was more active for ΔV_{self} compared to ΔV_{other} in self-interested participants, who care more about their own profits than the other's loss. Conversely, for other-regarding participants, those who place higher weight on ΔV_{other} compared to ΔV_{self}, the vmPFC was more active for ΔV_{other} compared to ΔV_{self} (r(25)=0.49, p=0.009). Thus, the vmPFC represents in each participant her own idiosyncratic *balance* between value to self and value to other. This finding is consistent with the role of the vmPFC as an integrator of various attributes of choice to one single value.

Discussion

Using a novel modification of the Message task, we found that both self-interest and regard for others contribute independently to (dis)honesty. Crucially, our results suggest that these motives rely on distinct neural processes, with self-interest involving the lateral prefrontal and parietal cortices, and regard for others the right temporoparietal junction and the valuation system (vmPFC and VS). Furthermore, we find a combination of motives in the vmPFC, consistent with its role as an integrator of value. That is, in participants who care more about their own profit compared to the other's loss, the vmPFC was more sensitive to payoffs to self than losses to other; conversely, in participants who care more about the other compared to themselves, the vmPFC showed higher activity for their partner's potential losses than to their own potential gains.

Pilot studies conducted in our lab ensured that 100% of Receivers indeed choose according to the message sent by the Sender, and 100% of Senders indicated that they believed that the Receiver would choose the option they recommended. This reassured us that the Senders' decisions are purely honesty-related, and not strategic. Previous research has empirically demonstrated that increasing the consequences of the lie decreases its occurrence (Gneezy, 2005). We have extended this notion to elucidate individual differences in this sensitivity. Furthermore, due to our within-participant design and orthogonality of the regressors, we captured independent self- and other-regarding motives. This independence coincides with the observation that some people would not lie, even if lying would *help* the Receiver (Erat and Gneezy, 2012), suggesting that for some people, the self-regarding motive is the only motive driving behavior. Thus, in some cases, honesty could be unrelated to the consequence of lying. Although our design does not include lies that help others (i.e., altruistic white lies), we do observe participants with non-significant weights on either self- or other-regarding motives. Thus, some participants' behavior can be described using a single motive.

In the context of choices, differences in reaction times should be interpreted carefully. Reaction times could attest to the ease of choice (Milosavljevic et al., 2010; Krajbich et al., 2015), or to the rate in which evidence accumulates in favor of one of the alternatives (Krajbich et al., 2012), which is related to the strength-of-preferences (Krajbich et al., 2015). Although all participants experienced objectively the same choice set, their variation in preferences yielded considerable variations in decision times for different choices. Along the same lines as previous findings involving prosocial decisions (Krajbich et al., 2015), we find that honest participants are quicker to tell the truth, whereas dishonest participants are quicker to lie. In other words, acting out of character takes longer. This result goes against dual-process models, which assume similar preferences for all participants (e.g., everyone is selfish at heart, and need extra time to overcome their selfish urge to act prosocially).

In accordance with previous research (Kozel et al., 2004; Christ et al., 2009; Sip et al., 2010), we find that the mPFC, dlPFC, and insula are more active during lying compared to telling the truth. The opposite contrast yielded activity in the TPJ and STS. This somewhat diverges from previous studies, which often report null result for a contrast of truth-telling compared to lying (Spence, 2004). One potential explanation for this discrepancy is that in our design, telling the truth should not be considered a "baseline" response. Pitting monetary gain with honesty means that telling the truth necessarily results in forfeiting some amount of money.

We find that the LPFC represents potential profits from lying. Recently, Crockett et al. (2017), demonstrated that the LPFC negatively represents potential ill-gotten gains. The authors asked participants to inflict pain on others and on themselves for profit and found that individual differences in harm-aversion correlated with the activity that represents monetary gain in the LPFC. Specifically, activity in this region decreased as the amount of money that could potentially be gained by physically hurting another individual grew. The authors attributed this finding to a sense of blame, showing that the LPFC encodes the level of blameworthiness – higher gains are associated with lower blame and thus lower activity in the LPFC. Our results are in line with this notion, showing a negative relationship between potential monetary profit through dishonesty and LPFC activity, extending them to a new, previously undiscussed, domain of moral decision-making.

Another explanation for the involvement of the lateral prefrontal cortex is related to its perhaps most known role – cognitive control (Badre and Nee, 2018). We examined our findings in this light, by testing the behavioral relationship between value to self and reaction times. The logic behind it being that smaller potential profits may require more cognitive control, which would be reflected in longer reaction times. We do not find any direct evidence for such an effect. However, we do find that generally honest participants (based on their overall dishonesty scores) take longer to choose when they can gain a large profit from lying, while generally dishonest participants take longer to choose when the potential profit from lying is small. Thus, levels of conflict seem to be related to idiosyncratic preferences. These results also indicate how misleading group averages may be, and highlight the importance of accounting for individual differences when interpreting neural findings.

We find that the right TPJ negatively tracks the value for other (ΔV_{other}). That is, after accounting for individual differences in regard for others, we find that the rTPJ encodes the Receiver’s loss positively in participants who care more about it (high on regard for others). Thus, the sensitivity of the rTPJ to consequences of dishonesty is likely to drive some individuals to act prosocially and others selfishly. Interestingly, this area has been implicated in deception, but it did not show up in the overlap between deception-related and executive-control-related brain maps (Christ et al., 2009). Thus, the TPJ may represent a different aspect of dishonesty. We propose that this aspect is its social consequences. The TPJ is a known major hub in the social cognition neural network. It is selectively activated when interacting with social agents (Carter et al., 2012), and is thought to be responsible for the ability to represent other people's minds, a process sometimes called perspective taking or mentalizing (Saxe and Kanwisher, 2003; Samson et al., 2004; Schurz et al., 2014). The locus of our activation corresponds to the posterior rTPJ, an area highly connected with the vmPFC and implicated in a wide array of mentalizing tasks (Schurz et al., 2014). In the moral domain, the rTPJ is also involved in forming judgments about others' moral acts (Young et al., 2007, 2010). Here, we provide the first evidence that the potential outcomes of moral acts (namely, dishonesty) are represented by the rTPJ, as well. Moreover, we show that the individual's sensitivity to these potential outcomes can explain their prosocial behavior.

In addition to the rTPJ, we find preference-modulated value representation for other's losses in the vmPFC and VS. Previous research has linked the vmPFC to social concepts like fairness and guilt. For example, making inequitable choices deactivates the vmPFC (Zaki and Mitchell, 2011), irrespective of who gains from this inequity – self or other. Lesions to the vmPFC reduce prosociality and trustworthiness, as measured by the Dictator, Ultimatum and Trust games, and vmPFC patients exhibit an overall diminished sensitivity to guilt (Krajbich et al., 2009). Dishonest choices that result in large losses to another should elicit guilt; therefore, it seems plausible, that motive-modulated representation of other's value would take place in the vmPFC. In this view, the sensitivity of the valuation system to others' wellbeing is linked to actual prosocial or antisocial choices the individual makes.

Finally, we demonstrate that the vmPFC also represents the *integrative value* of self and other (how much the Sender would gain from a dishonest message vs. how much the Receiver would gain from an honest one). Importantly, this integration process in the vmPFC occurs only when accounting for the weight each individual places on own and other's value, implying that the nature of the representation is subjective – personal preference modulate the objective amounts of money presented on the screen. We find that if an individual cares more about themselves than others, her valuation system will be more sensitive to her own profits. Alternatively, the valuation system of someone who cares more about others than herself, would be more sensitive to others' profits. This balanced representation of value suggests a neural instantiation of observed differences in behavior. Our findings are in line with abundant research on the role of the vmPFC in value computation and representation (Levy and Glimcher, 2012; Bartra et al., 2013; Clithero and Rangel, 2013; Ruff and Fehr, 2014), suggesting that when two values or motives conflict, the vmPFC represents a comparison of the two.

In summary, we found that dishonest behavior varies dramatically between individuals, both in which motive drives the behavior and to what extent. Moreover, we find that two distinct motives can be identified from behavior and traced back to separate neural representations. At the behavioral level, individual's levels of self-interest and regard for others contribute independently to a choice to lie. On the neural level, the LPFC and TPJ represent value to self and value to other, respectively. Importantly, self- and other-regarding motives for (dis)honesty affect this representation of value – in the LPFC, the TPJ, and the valuation system. These findings suggest a neural instantiation of individual differences in behavior; while prosocial individuals' neural valuation systems are more sensitive to others' wellbeing, those of selfish individuals are more sensitive to their own.

Method

Participants

Thirty-three participants enrolled in the study (22 females; age *M*=25.35, 19-30). Participants gave informed written consent before participating in the study, which was approved by the local ethics committee at Tel-Aviv University. All participants were right-handed, and had a normal or corrected-to-normal vision. Of the 33 participants, five were excluded from all analyses due to a lack of variability in their behavioral responses – they either lied on more than 90% (n=3) or on less than 10% (n=2) of the trials. An additional participant was excluded only from neural analyses due to excessive head movements during the scans (>3 mm). Participants were paid a participation fee and the amount of money they won on a randomly drawn and implemented trial.

Experiment

Task

On each trial, the participant lying in the fMRI scanner (*Sender*) was asked to send the following message to her partner (*Receiver*): "Option __ is most profitable for you". Senders watched a screen with 4 'doors'. Two doors were empty (indicating payoffs of $0 for both participants), and two doors contained non-zero payoff information for themselves and for their partner (the Receiver; see Fig. 1). Above the doors appeared the message text. Participants were given 6 seconds to indicate their choice by pressing one of four buttons on a response box. After choosing, the chosen door was highlighted and its number appeared in the text (e.g., "Option 3 is most profitable for you", if a participant chose door number 3). The duration of this decision screen was set to have a minimum of 1.5 s and to make up for a total trial duration of 7.5 s. For example, if a participant made a choice after 3.5 seconds, the decision screen appeared for 4 seconds. If no choice was made in the allotted time, a “no choice” feedback screen appeared for 1.5 s. Afterwards, a fixation screen appeared for 6-10.5 seconds.

Payoff structure

The payoffs were intended to create a conflict between honesty and monetary profit. On each trial, each door contained some amount of money for the Sender and some amount of money for the Receiver. A truthful message (*Truth* option) is defined as when the Sender is choosing the door (the message to send) that results in a larger amount of money for the Receiver than if the other door was chosen (the door that would have resulted in a deceptive message (*Lie*)). A deceptive message (*Lie* option) is defined as when the Sender is choosing the door that results in a smaller amount of money for the Receiver (and a larger amount of money for the Sender) compared to the *Truth* option. Payoffs varied on a trial-by-trial basis, ranging between 10-42₪ (1₪ ≈ 0.3USD) for the Sender, and 1-31₪ for the Receiver. Critically, the potential profits for the Sender from lying (ΔV_{self}, $\$Self_{Lie}$ -$\$Self_{Truth}$; range 0-12₪) varied independently from the potential losses for the Receiver (ΔV_{other}, $\$Other_{Truth}$ -$\$Other_{Lie}$; range 1-20₪; r(38) = 0.25, p = 0.11). Figure 1 depicts an example trial, in which choosing to tell the truth would result in 10₪ for the Sender and 9₪ for the Receiver, whereas lying would result in 15₪ for the Sender and only 3₪ for the Receiver. In this case, the potential profit for the Sender (ΔV_{self}) is 5₪ $(15 - 10)$, and the potential loss to the Receiver (ΔV_{other}) is 6₪ $(9 - 3)$. Importantly, to avoid any interfering motives (e.g. envy), all trials of interest (38 out of 40) had a higher payoff for the Sender than for the Receiver. That is, in these trials the Sender was always better off than the Receiver, irrespective of the chosen door. Two additional trials were *catch trials*, offering a 0₪ payoff for the Sender from lying (i.e., $\$Self_{Lie}$=$\$Self_{Truth}$). For example, the Truth option could hold 23₪ for the

Receiver and 38₪ for the Sender, whereas a Lie option would hold only 15₪ for the Receiver, and the same 38₪ for the Sender. In this case, there is no conflict between honesty and gain for the Sender, and these trials served merely to ensure the participants are attentive. A full list of the payoffs appears in table S1 and Figure S1.

A key component of the task is the addition of two empty doors to each trial, containing zero money for each player (doors 1 & 3 in the Figure 1 example). The order of the doors was randomized across trials. These options ensured that the message the Sender sends is indeed followed by the Receiver, because deviating from the recommended door may result in opening an empty door. Consider the example in Figure 1: if the Sender chooses to lie, she sends a message regarding door #4. All the Receiver will see is that door #4 is highlighted. The Receiver does not know how much money is behind which door, but she does know that two of the doors have no money at all and that the Sender wants to gain at least some amount of money. Even if the Receiver believes she is lied to, it is in her best interest to choose to open door #4, otherwise she (and the Sender) face 66% chance of winning no money at all.

Procedure

Participants arrived to the Imaging Center and met the experimenter and a confederate acting as the Receiver. The confederate was always a Caucasian female, aged around 25, to avoid any influence of social factors on the Sender's decision-making. Both the Sender and Receiver read written instructions, signed consent forms and underwent a training stage of the task – playing as both the Sender and the Receiver. Training as Sender was intended to familiarize the participant with the Message Task. Training on the Receiver's role allowed them to experience what happens when the Receiver does not follow the recommended Message, and ensure they understood the consequences of sending a truthful or deceptive message.

Each participant completed four scans. Forty unique payoff trials were randomly interspersed in a given scan, making up 160 trials per participant (40 unique payoffs × 4 repetitions). At the end of the experiment, one trial was selected randomly and presented to the Sender as the message that will be sent to the Receiver. Pilot studies in our lab, using real particiapnts acting as Receiver's revealed that 100% of them chose according to the message sent by the Sender. Therefore, in the current study, we automatically chose the Receiver's choice to always be according to the Sender's message, and paid the Sender whichever amount was associated with that option.

After the scan, participants completed a short debriefing questionnaire. Debriefing consisted of several demographic questions and questions regarding the choices the participant made in the task. Specifically, we aimed to ensure participants were not suspicious of the confederate, by asking how the identity of the Receiver affected their choices.

Image acquisition and processing

Scanning was performed at the Strauss Neuroimaging Center at Tel Aviv University, using a 3T Siemens Prisma scanner with a 64-channel Siemens head coil. Anatomical images were acquired using MPRAGE, which comprised 208 1-mm thick axial slices at an orientation of -30° to the AC–PC plane. To measure blood oxygen level-dependent (BOLD) changes in brain activity task performance, a T2*-weighted functional multi-band EPI pulse sequence was used (TR = 1.5s; TE = 30 ms; flip angle = 70°; °; matrix =

86 × 86; field of view (FOV) = 215 mm; slice thickness = 2.5 mm). 50 axial (−30° tilt) slices with no inter-slice gap were acquired in ascending interleaved order.

BrainVoyager QX (Brain Innovation) was used for image analysis, with additional analyses performed in Matlab. Functional images were sinc-interpolated in time to adjust for staggered slice acquisition, corrected for any head movement by realigning all volumes to the first volume of the scanning session using six-parameter rigid body transformations, and de-trended and high-pass filtered to remove low-frequency drift in the fMRI signal. Data were then spatially smoothed with a Gaussian kernel of 5 mm (full-width at half-maximum), co-registered with each participant's high-resolution anatomical scan and normalized using the Montreal Neurological Institute (MNI) template. All spatial transformations of the functional data used trilinear interpolation.

Behavioral analysis

Overall deception & reaction times

First, we removed no-response trials (*M*=0.9%, *SD*=1.2% of trials), and trials in which participants chose an empty door (option of $0 to both players; *M*=1.4%, *SD*=1.3% of trials). Finally, we removed the catch trials (trials with identical payoffs to the Sender for Truth and Lie; *M*=4.8%, *SD*=0.03% of trials). All further analyses were performed on this subset of trials (*M*=92.7%, *SD*=3.2% of trials, range: 137-152 trials). We defined overall dishonesty rates per participant as the number of trials they lied out of this total number of trials.

To examine how potential profits from lying affect decision time, we conducted two analyses. First, we regressed value to self (ΔV_{self}) onto reaction times, clustering the errors per participant, to get an across-participant measure of the relationship between the two variables. Second, we ran participant-specific correlations of reaction times and value to self, yielding a correlation coefficient per participant. Then, we examined whether the resulting correlation is related to other behavioral measures. To do so, we used each participant's correlation coefficient (r-value) calculated from the first correlation analysis, and correlated it with the overall dishonesty and with self-regarding motive for dishonesty.

Analysis of motives

To estimate the contribution of self- and other-regarding motives to dishonest behavior, we fitted a linear regression per participant. For each unique payoff, we calculated each subject's probability to lie by averaging her choices across the four repetitions. The probability to lie served as the dependent variable. The independent variables were the normalized profit to self and loss to other (ΔV_{self} & $\Delta V_{\mathrm{other}}$, respectively):

$$probability(lie) = \beta_0 + \beta_1 \widetilde{\Delta V}_{self} + \beta_2 \widetilde{\Delta V}_{other} + \varepsilon$$

where probability to lie ranges from 0 to 1, and ~ represents the z-transformed profits. We refer to the estimated coefficients as *self-interest* ($\beta\Delta V_{\mathrm{self}}$) and *regard for others* ($\beta\Delta V_{\mathrm{other}}$).

Statistical analyses

All reported t-tests are two-tailed. All reported correlations are Pearson correlations.

fMRI analysis

Statistical significance

Unless specified otherwise, we used cluster-size threshold for multiple comparison correction. Cluster-defining threshold was set to 0.005, with a 1000 Monte Carlo simulations to achieve family-wise error of 0.05.

Value representation for self and other

To identify neural correlates of value to self (ΔV_{self}) and value to other (ΔV_{other}), we constructed a general linear model (GLM1) with the following predictors: (1) options period – a box-car function of the duration from trial onset until the participant made a decision; (2) decision period – a box-car function of the duration from decision until ITI; (3) ΔV_{self} – the difference between the profit (in ₪) to Sender in the Lie option and the Truth option (i.e., gain from lying; $\$\text{Self}_{\text{Lie}}$ -$\$\text{Self}_{\text{Truth}}$); (4) ΔV_{other} – the difference between profit (in ₪) to Receiver in the Truth option and the Lie option ($\$\text{Other}_{\text{Truth}}$ - $\$\text{Other}_{\text{Lie}}$). Additional nuisance predictors included six motion-correction parameters and a mean signal from the ventricles, accounting for respiration.

Neural correlates of Individual differences

Using GLM1, we examined whether behavioral individual differences, in the motives driving dishonest behavior (i.e., $\beta\Delta V_{\text{self}}$ and $\beta\Delta V_{\text{other}}$), can explain variation in the neural representation of value. We therefore conducted two analyses: First, we looked for voxels exhibiting a significant correlation between neural representation of ΔV_{self} and $\beta\Delta V_{\text{self}}$ – the behavioral measure of self-interest. That is, first we identified for each subject a neural activation map representing ΔV_{self}. Thereafter, we regressed the behavioral estimate of value to self ($\beta\Delta V_{\text{self}}$) across participants onto the activation map identified in the first stage. We repeated the analysis with ΔV_{other} and $\beta\Delta V_{\text{other}}$. This analysis yielded two voxel-wise correlation maps (one correlating BOLD ΔV_{self} and behavioral $\beta\Delta V_{\text{self}}$ and one correlating BOLD ΔV_{other} and behavioral $\beta\Delta V_{\text{other}}$), as implemented in BrainVoyager.

Chosen value representation

To uncover voxels representing the chosen value of each trial, we first constructed a second GLM (GLM2) with the following predictors: (1) options period – a box-car function of the duration from trial onset until participant made a decision; (2) decision period – a box-car function of the duration from decision until ITI; (3) chosen value – the amount of money (in ₪) that the Sender will receive in the chosen door; (4) unchosen value – the amount of money (in ₪) that the Sender would have received in the unchosen door; (5) Truth – an indicator function for trials in which the Sender chose the Truth option; and (6) Lie – an indicator function for trials in which the Sender chose the Lie option. Additional nuisance predictors included six motion-correction parameters and a mean signal from the ventricles, accounting for respiration. To reveal voxels representing chosen value, we contrasted chosen value with unchosen value (predictors 3 and 4). We restricted the analysis to the valuation system only – the vmPFC and ventral striatum, defined using the results of a meta-analysis (Bartra et al., 2013). We used FDR multiple-comparison correction at a p=0.05 level.

Neural correlates of deception

We used GLM2 to contrast Truth (trials in which the participant sent a truthful message) and Lie trials (in which the participant sent a deceptive message). The “chosen value” predictor controlled for the amount of money participants stood to gain on each trial, to ensure that the observed patterns of neural response reflected solely the choice to lie or tell the truth, irrespective of reward size.

Data Availability

All statistical maps and computer code used to analyze the fMRI data are available on OSF.org (https://osf.io/bvuxc/).

References

Abe N (2009) The neurobiology of deception: evidence from neuroimaging and loss-of-function studies. Current opinion in neurology 22:594–600.

Abe N (2011) How the Brain Shapes Deception: An Integrated Review of the Literature. The Neuroscientist 17:560–574.

Andreoni J, Miller J (2002) Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. Econometrica 70:737–753.

Badre D, Nee DE (2018) Frontal Cortex and the Hierarchical Control of Behavior. Trends in Cognitive Sciences 22:170–188.

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

Bhatt S, Mbwana J, Adeyemo A, Sawyer A, Hailu A, VanMeter J (2009) Lying about facial recognition: An fMRI study. Brain and Cognition 69:382–390.

Carter RM, Bowling DL, Reeck C, Huettel SA (2012) A Distinct Role of the Temporal-Parietal Junction in Predicting Socially Guided Decisions. Science 337:109–111.

Christ SE, Essen DCV, Watson JM, Brubaker LE, Mcdermott KB (2009) The Contributions of Prefrontal Cortex and Executive Control to Deception : Evidence from Activation Likelihood Estimate Meta- analyses. :1557–1566.

Clithero J a, Rangel A (2013) Informatic parcellation of the network involved in the computation of subjective value. Social Cognitive and Affective Neuroscience Available at: http://scan.oxfordjournals.org/content/early/2013/08/21/scan.nst106.abstract%5Cnhttp://scan.oxfordjournals.org/content/early/2013/07/24/scan.nst106.abstract.

Crockett MJ, Siegel JZ, Kurth-Nelson Z, Dayan P, Dolan RJ (2017) Moral transgressions corrupt neural representations of value. Nature Neuroscience 20:879–885.

Erat S, Gneezy U (2012) White Lies. Management Science 58:723–733.

Fehr E, Camerer CF (2007) Social neuroeconomics: the neural circuitry of social preferences. Trends in Cognitive Sciences 11:419–427.

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

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

Greene JD, Paxton JM (2009) Patterns of neural activity associated with honest and dishonest moral decisions. Proceedings of the National Academy of Sciences of the United States of America 106:12506–12511.

Hare T a, Camerer CF, Knoepfle DT, Rangel A (2010) Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. The Journal of neuroscience : the official journal of the Society for Neuroscience 30:583–590.

Jacobsen C, Fosgaard TR, Pascual-Ezama D (2018) Why do we lie? A practical guide to the dishonesy literature. Journal of Economic Surveys 32:357–387.

Janowski V, Camerer C, Rangel A (2013) Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. Social Cognitive and Affective Neuroscience 8:201–208.

Jenkins AC, Zhu L, Hsu M (2016) Cognitive neuroscience of honesty and deception: a signaling framework. Current Opinion in Behavioral Sciences 11:130–137.

Kozel AF, Revell LJ, Lorberbaum JP, Shastri A, Elhai JD, Horner MD, Smith A, Nahas Z, Bohning DE, George MS (2004) A Pilot Study of Functional Magnetic Resonance Imaging Brain Correlates of Deception in Healthy Young Men. Journal Of Neuropsychiatry 16:295–305.

Kozel FA, Johnson KA, Mu Q, Grenesko EL, Laken SJ, George MS (2005) Detecting Deception Using Functional Magnetic Resonance Imaging.

Krajbich I, Adolphs R, Tranel D, Denburg NL, Camerer CF (2009) Economic Games Quantify Diminished Sense of Guilt in Patients with Damage to the Prefrontal Cortex. Journal of Neuroscience 29:2188–2192.

Krajbich I, Bartling B, Hare T, Fehr E (2015) Rethinking fast and slow based on a critique of reaction-time reverse inference. Nature Communications 6:1–9.

Krajbich I, Lu D, Camerer C, Rangel A (2012) The Attentional Drift-Diffusion Model Extends to Simple Purchasing Decisions. Frontiers in Psychology 3 Available at: http://journal.frontiersin.org/article/10.3389/fpsyg.2012.00193/abstract [Accessed December 5, 2018].

Langleben DD, Loughead JW, Bilker WB, Ruparel K, Childress AR, Busch SI, Gur RC (2005) Telling truth from lie in individual subjects with fast event-related fMRI. Human Brain Mapping 26:262–272.

Levy DJ, Glimcher PW (2012) The root of all value: a neural common currency for choice. Current Opinion in Neurobiology:1–12.

Liang C-Y, Xu Z-Y, Mei W, Wang L-L, Xue L, Lu DJ, Zhao H (2012) Neural correlates of feigned memory impairment are distinguishable from answering randomly and answering incorrectly: An fMRI and behavioral study. Brain and Cognition 79:70–77.

Lisofsky N, Kazzer P, Heekeren HR, Prehn K (2014) Investigating socio-cognitive processes in deception: A quantitative meta-analysis of neuroimaging studies. Neuropsychologia 61:113–122.

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.

Milosavljevic M, Malmaud J, Huth A, Koch C, Rangel A (2010) The Drift Diffusion Model Can Account for the Accuracy and Reaction Time of Value-Based Choices Under High and Low Time Pressure. SSRN Electronic Journal Available at: http://www.ssrn.com/abstract=1901533 [Accessed December 5, 2018].

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

Morelli SA, Sacchet MD, Zaki J (2015) Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. NeuroImage 112:244–253.

Rilling JK, Gutman DA, Zeh TR, Pagnoni G, Berns GS, Kilts CD (2002) A Neural Basis for Social Cooperation. Neuron 35:395–405.

Ruff CC, Fehr E (2014) The neurobiology of rewards and values in social decision making. Nature reviews Neuroscience 15:549–562.

Samson D, Apperly IA, Chiavarino C, Humphreys GW (2004) Left temporoparietal junction is necessary for representing someone else's belief. Nature Neuroscience 7:499–500.

Saxe R, Kanwisher N (2003) People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind." NeuroImage 19:1835–1842.

Schurz M, Radua J, Aichhorn M, Richlan F, Perner J (2014) Fractionating theory of mind: A meta-analysis of functional brain imaging studies. Neuroscience & Biobehavioral Reviews 42:9–34.

Sip KE, Lynge M, Wallentin M, McGregor WB, Frith CD, Roepstorff A (2010) The production and detection of deception in an interactive game. Neuropsychologia 48:3619–3626.

Smith DV, Clithero JA, Boltuck SE, Huettel SA (2013) Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. Social Cognitive and Affective Neuroscience 9:2017–2025.

Spence SA (2004) The deceptive brain. Journal of the Royal Society of Medicine 97:6–9.

Spence SA, Farrow TFD, Herford AE, Wilkinson ID, Zheng Y, Woodruff PWR (2001) Behavioural and functional anatomical correlates of deception in humans: Neuroreport 12:2849–2853.

Volz KG, Vogeley K, Tittgemeyer M, von Cramon DY, Sutter M (2015) The neural basis of deception in strategic interactions. Frontiers in Behavioral Neuroscience 9 Available at: http://journal.frontiersin.org/Article/10.3389/fnbeh.2015.00027/abstract [Accessed December 5, 2018].

Vrij A (2004) Detecting lies and deceit: the psychology of lying and the implications for professional practice. Chichester: Wiley.

Yang Y, Raine A, Lencz T, Bihrle S, Lacasse L, Colletti P (2005) Prefrontal white matter in pathological liars. British Journal of Psychiatry 187:320–325.

Yang Y, Raine A, Narr KL, Lencz T, LaCasse L, Colletti P, Toga AW (2007) Localisation of increased prefrontal white matter in pathological liars. British Journal of Psychiatry 190:174–175.

Yin L, Reuter M, Weber B (2016) Let the man choose what to do: Neural correlates of spontaneous lying and truth-telling. Brain and Cognition 102:13–25.

Young L, Camprodon JA, Hauser M, Pascual-Leone A, Saxe R (2010) Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. Proceedings of the National Academy of Sciences 107:6753–6758.

Young L, Cushman F, Hauser M, Saxe R (2007) The neural basis of the interaction between theory of mind and moral judgment. Proceedings of the National Academy of Sciences 104:8235–8240.

Zaki J, Lopez G, Mitchell JP (2014) Activity in ventromedial prefrontal cortex co-varies with revealed social preferences: evidence for person-invariant value. Social Cognitive and Affective Neuroscience 9:464–469.

Zaki J, Mitchell JP (2011) Equitable decision making is associated with neural markers of intrinsic value. Proceedings of the National Academy of Sciences 108:19761–19766.

Zhu L, Jenkins AC, Set E, Scabini D, Knight RT, Chiu PH, King-Casas B, Hsu M (2014) Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. Nature neuroscience 17:1319–1321.

Figures

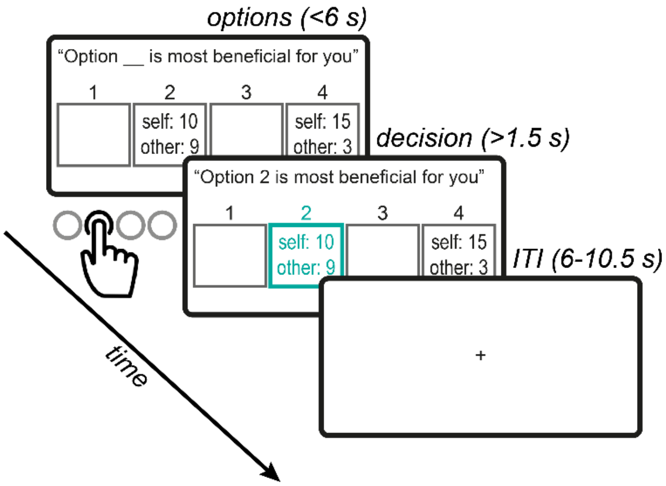

Figure 1. Trial timeline. On each trial, the participant (a "Sender") chose which message to send to the Receiver, out of four options. The text at the top of the screen is the message that would be sent on this trial to the Receiver. Four options are revealed, each one consisting of some amount of money for the Sender ("self") and some for the Receiver ("other"). One option was always truthful (the one more beneficial for the Receiver; #2 in the example) and one deceptive (#4). Payoffs to both players and locations varied between trials. The Sender had 6 seconds to indicate her choice, after which the chosen option was highlighted and stayed on the screen for the remainder of the trial.

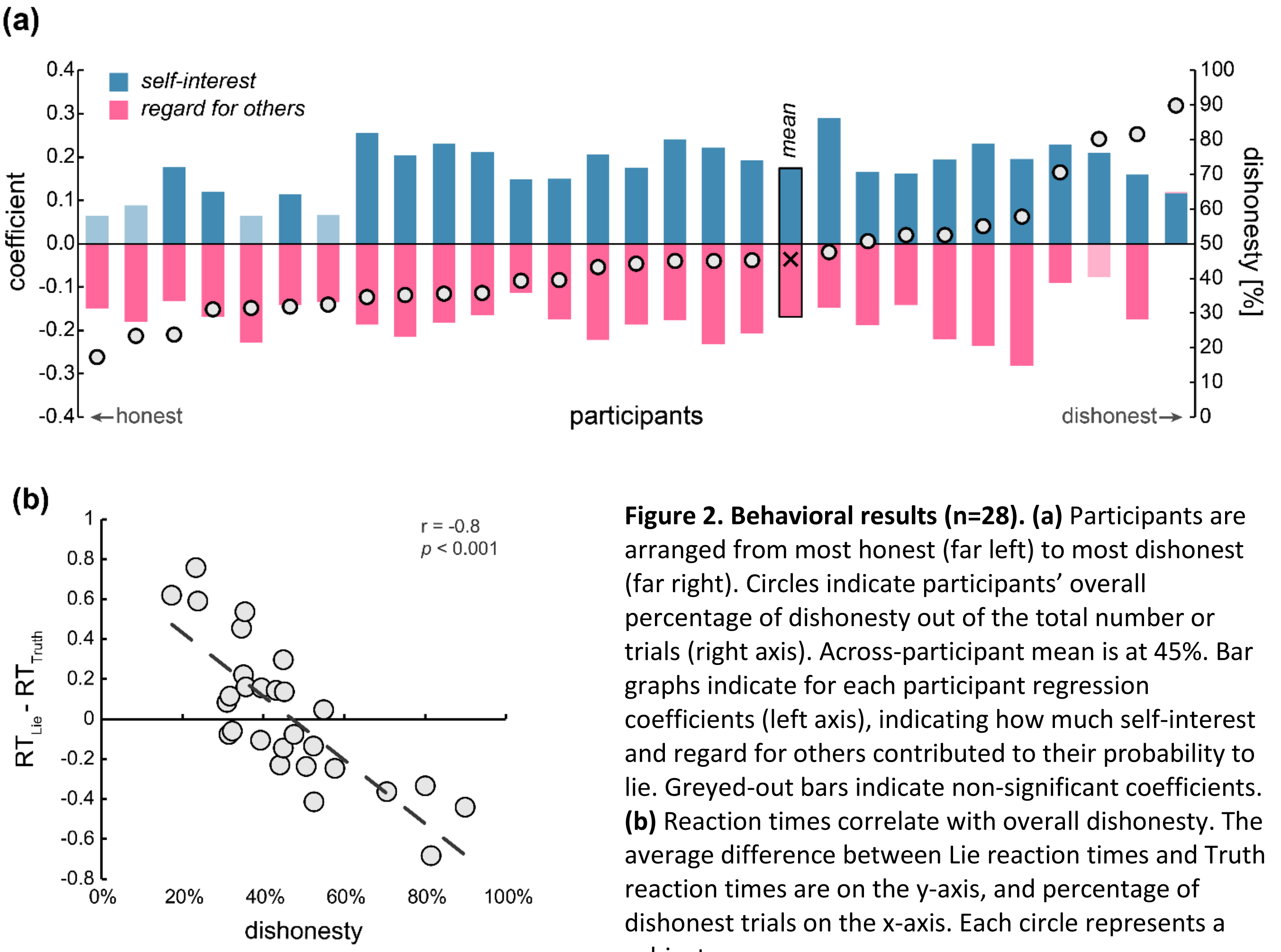

Figure 2. Behavioral results (n=28). (a) Participants are arranged from most honest (far left) to most dishonest (far right). Circles indicate participants' overall percentage of dishonesty out of the total number or trials (right axis). Across-participant mean is at 45%. Bar graphs indicate for each participant regression coefficients (left axis), indicating how much self-interest and regard for others contributed to their probability to lie. Greyed-out bars indicate non-significant coefficients. **(b)** Reaction times correlate with overall dishonesty. The average difference between Lie reaction times and Truth reaction times are on the y-axis, and percentage of dishonest trials on the x-axis. Each circle represents a subject.

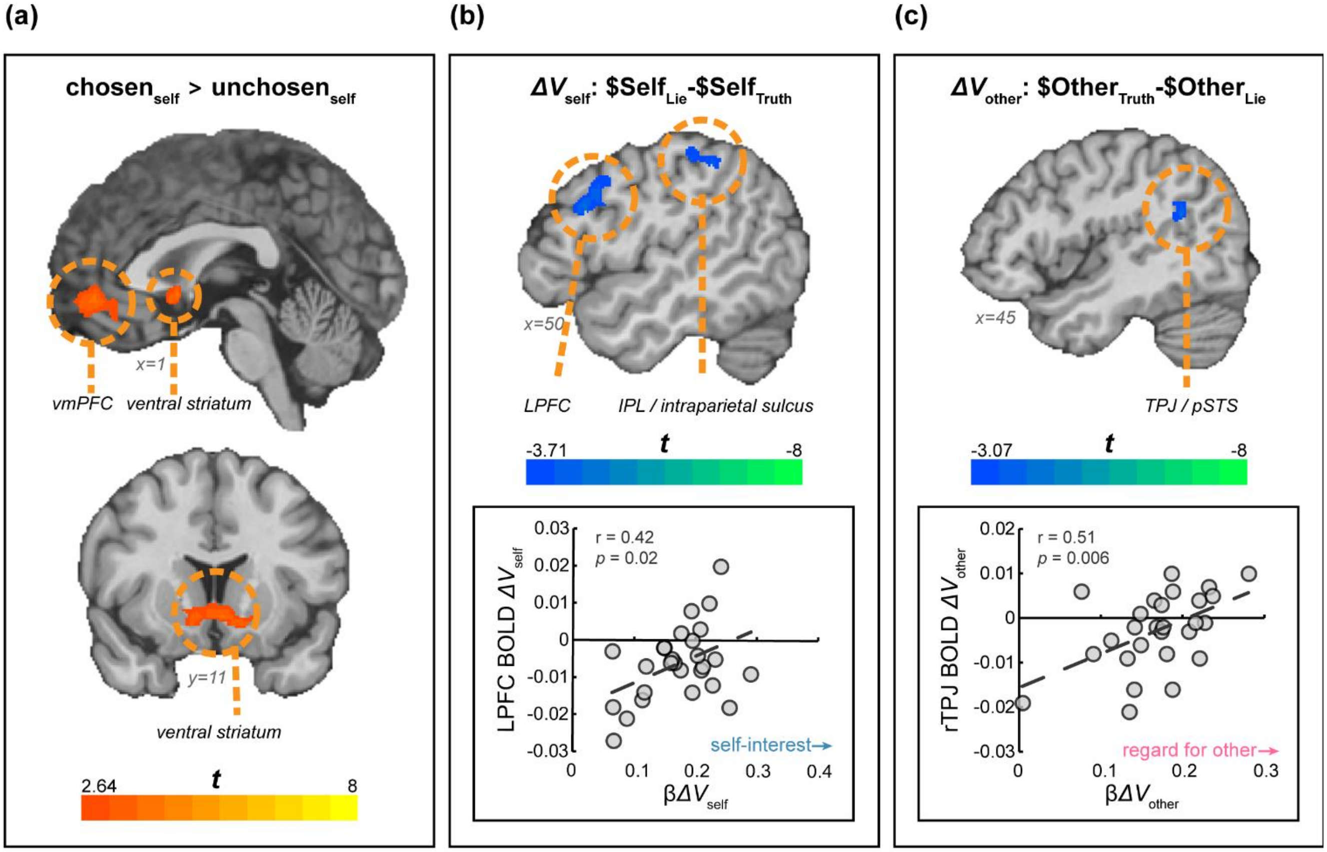

Figure 3. Neural representation of values (n=27). (a) Amount of money for the Sender in the chosen option vs. the unchosen option. The activation map was masked using value related ROIs taken from a meta-analysis (Bartra et al., 2013). Map at *q*(FDR)=0.05. **(b)** Voxels sensitive to Sender's potential profits from dishonesty (value for self; ΔV_{self}). Map thresholded at *p*=0.001, cluster-size corrected (top). BOLD response in the LPFC is positively correlated with the behavioral measure of self-interest (bottom). **(c)** Voxels sensitive to Receiver's potential losses from dishonesty (value for other; ΔV_{other}). Map thresholded at *p*=0.005, cluster-size corrected (top). BOLD response in the rTPJ is positively correlated with the behavioral measure of regard for others.

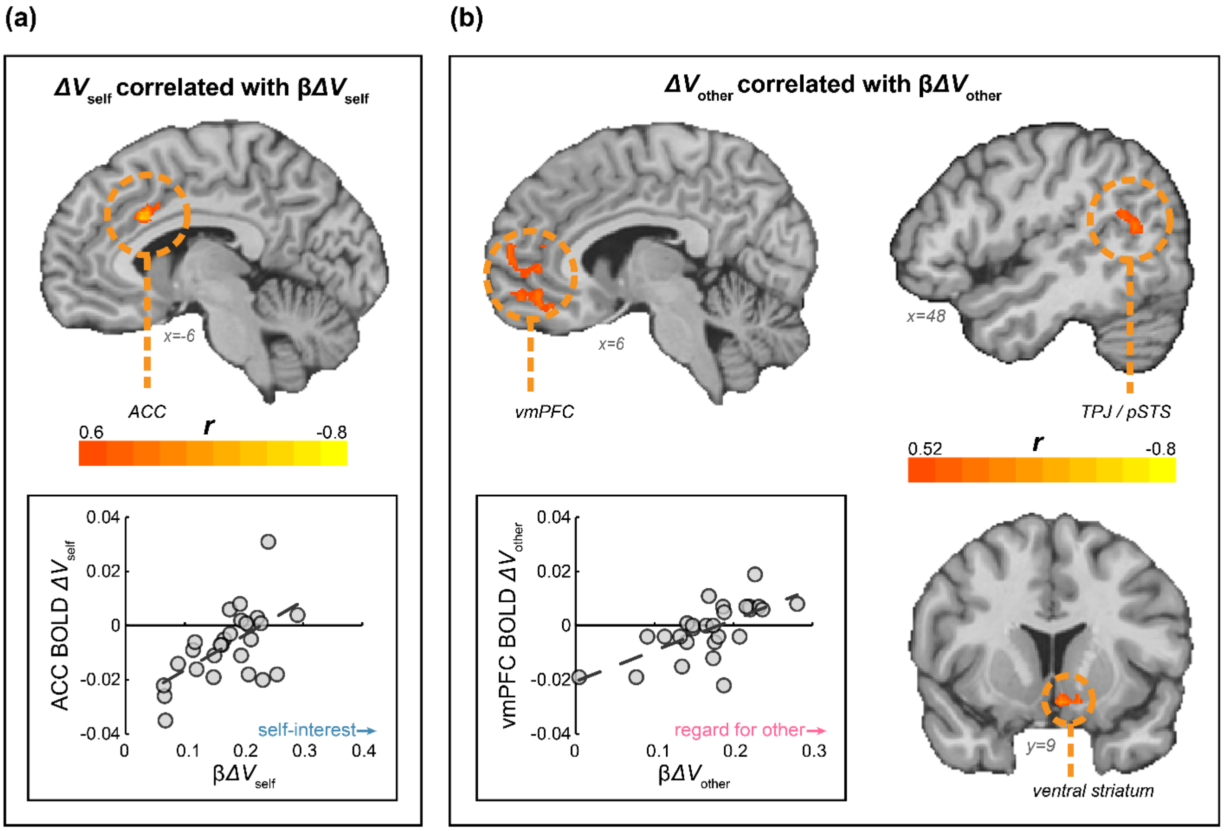

Figure 4. A whole-brain analysis of Individual differences in behavior correlated with neural differences in value representation (n=27). (a) The ACC represents preference-modulated value for self. Participants high on self-interest have higher self-value representation in the ACC. Map thresholded at *p*=0.001, cluster-size corrected. **(b)** vmPFC, ventral striatum, and rTPJ represent preference-modulated value for other. Participants with high regard for others have higher other-value representation in these areas. Map thresholded at *p*=0.005, cluster-size corrected.

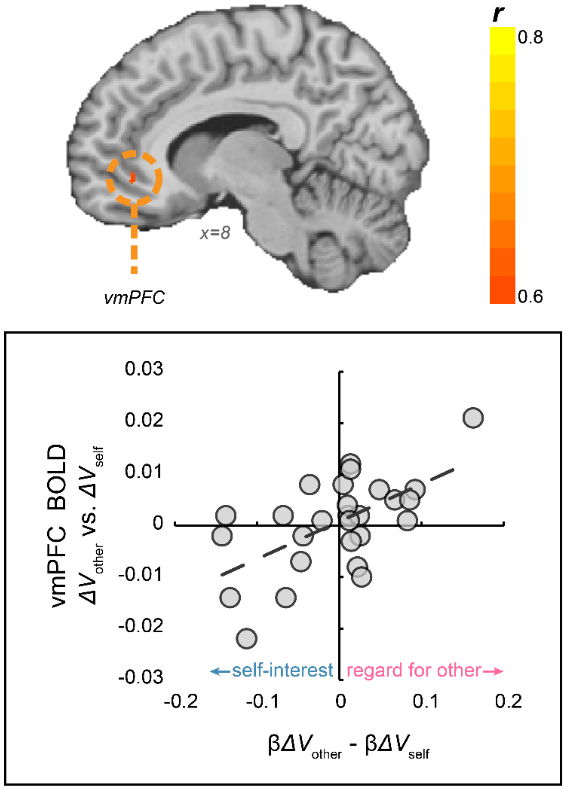

Figure S4. Value to self vs. value to other, modulated by preferences. Contrasting potential profit for Sender (value to self) with potential losses to Receiver (value to other), modulated by individual differences in the balance between regard for other and self-interest. Map thresholded at *p*=0.001. Brain masked with an ROI defined based on a meta-analysis of the valuation system (Bartra et al., 2013). N=28.

Tables

		MNI coordinate				cluster
contrast	**brain area**	***x***	***y***	***z***	***T***	**size**
Truth vs. Lie	medial PFC	-3	29	38	4.06	128
	(left) dorsolateral PFC	-47	31	28	3.93	23
	(left) insula	-34	23	-2	4	31
	(right) insula	30	25	-2	4.37	29
	(right) TPJ	64	-28	-25	-4.02	116
	(left) superior temporal sulcus	-58	-18	-5	-4.09	43
	cerebellum	-9	-65	-5	-4.44	81
value to self	(left) lateral PFC	-44	16	31	-3.53	409
(ΔV_{self})	(right) IPL	36	-50	47	-3.62	474
	(left) IPL	-26	-54	39	-3.6	1471
	pre-SMA	-2	15	50	-3.42	51
	(left) occipital	-45	-67	4	-3.33	72
	(right) occipital	45	-77	-3	-3.4	168
	(medial) occipital	-15	-70	1	-3.57	666
	cerebellum	5	-64	-39	-3.35	63
	(right) amygdala	28	-4	-14	-3.5	46
value to other	(right) TPJ	45	-52	24	-3.37	33
(ΔV_{other})	precuneus	-7	-63	44	3.79	21
	(left) thalamus	-6	-18	11	3.37	26
	cuneus	-6	-92	22	3.36	32
	(left) occipital	-9	-73	3	-3.4	115
	(right) occipital	9	-73	-8	-3.42	108

Table 1. List of whole-brain contrast results. PFC: prefrontal cortex; IPL: inferior parietal cortex; TPJ: temporoparietal junction; SMA: supplementary motor area.

		MNI coordinate				cluster
contrast	**brain area**	***x***	***y***	***z***	***r***	**size**
value to self, modulated by self-interest	ACC	-3	14	35	0.57	119
	(left) occipitotemporal cortex	-42	-67	14	0.57	74
	(right) occipitotemporal cortex	43	-67	5	0.57	55
	(left) superior temporal cortex	-51	-23	3	0.54	47
	(right) temporal pole	-37	6	-41	0.56	98
	(right) insula & IFG	-45	4	-12	0.56	161
value to other, modulated by regard for others	(right) TPJ	50	-58	15	0.56	31
	medial PFC	5	53	8	0.56	69
	ventromedial PFC	7	47	-13	0.56	35
	ventral striatum	-5	7	-14	0.57	21

Table 2. List of whole-brain correlation results. ACC: anterior cingulate cortex; IFG: inferior frontal gyrus; TPJ: temporoparietal junction.