Cognitive control increases honesty in cheaters but cheating in those who are honest

Sebastian P. H. Speer[a,1], Ale Smidts[a], and Maarten A. S. Boksem[a]

[a]Rotterdam School of Management, Erasmus University, 3062 PA Rotterdam, The Netherlands

Edited by Joshua D. Greene, Harvard University, Cambridge, MA, and accepted by Editorial Board Member Michael S. Gazzaniga June 15, 2020 (received for review February 25, 2020)

Every day, we are faced with the conflict between the temptation to cheat for financial gains and maintaining a positive image of ourselves as being a "good person." While it has been proposed that cognitive control is needed to mediate this conflict between reward and our moral self-image, the exact role of cognitive control in (dis)honesty remains elusive. Here we identify this role, by investigating the neural mechanism underlying cheating. We developed a task which allows for inconspicuously measuring spontaneous cheating on a trial-by-trial basis in the MRI scanner. We found that activity in the nucleus accumbens promotes cheating, particularly for individuals who cheat a lot, while a network consisting of posterior cingulate cortex, temporoparietal junction, and medial prefrontal cortex promotes honesty, particularly in individuals who are generally honest. Finally, activity in areas associated with cognitive control (anterior cingulate cortex and inferior frontal gyrus) helped dishonest participants to be honest, whereas it enabled cheating for honest participants. Thus, our results suggest that cognitive control is not needed to be honest or dishonest per se but that it depends on an individual's moral default.

dishonesty | cognitive control | reward anticipation | self-referential thinking | fMRI

Imagine a friend sends you a link to a website where you can illegally stream recently released movies for free. Would you decide to stream the movie which you otherwise would have paid for? If so, how many movies would you stream? On a daily basis we are faced with the conflict between the temptation to violate moral standards to serve our self-interest and to uphold these moral standards, but how the brain resolves this conflict remains elusive.

When exposed to the opportunity to cheat, clearly, the (financial) rewards play a crucial role: the higher the reward, the more attractive the decision to cheat (1, 2). As the renowned British novelist Jonathan Gash so eloquently stated, "Fraud is the daughter of greed" (3). In line with this sentiment, behavioral research has demonstrated that greedy people find a variety of moral transgressions more acceptable and engage in such transgressions more often as compared to less greedy people (4). Indeed, neural responses in anticipation of reward, reflected in activity in the nucleus accumbens (Nacc), predict cheating behavior in a subsequent task (5). Collectively, these findings emphasize that higher rewards and stronger sensitivity to reward increase the likelihood of dishonesty.

Accumulating evidence from psychology, economics, and neuroscience has demonstrated, however, that people care about more than only maximizing their own monetary payoff, which is reflected in the high prevalence of prosocial behaviors such as altruism and reciprocity. People have internalized social norms and use these as an internal benchmark against which they compare their own behavior (6, 7). In the context of dishonesty, the way we view ourselves, our self-concept (8–10), may prevent us from cheating. People highly value honesty in others and also have strong beliefs in their own moral standards (11). Violating one's own moral standards will require a negative update of one's self-concept which is highly aversive (12). As a consequence, people are motivated to uphold their self-concept even if it comes at the cost of foregoing potential monetary gains (13). Hence, when given the opportunity to cheat, people are torn between the conflicting motivations to obtain desirable monetary gains versus the long-term goal of maintaining a positive self-image. Whereas the neural pattern associated with greed have been linked to cheating behavior (5), the neural processes linked to maintaining a positive self-image are yet to be discovered. Based on the behavioral research presented here, we hypothesize that self-referential thinking processes linked to a network of regions including the medial prefrontal cortex (MPFC), posterior cingulate cortex (PCC), and bilateral temporoparietal junctions (TPJs; 14–16) may underlie self-concept maintenance and promote honesty.

Several lines of research have proposed that cognitive control is needed to resolve the tension between reward and self-concept (5, 17–20). It has been found that this conflict is often settled with a compromise in which participants behave dishonestly enough to profit from the opportunity to cheat but honestly enough to maintain a positive self-image (13). While it is evident that cognitive control plays a crucial role in resolving this conflict, the precise nature of the role of cognitive control in moral decisions remains controversial (21).Two competing theories have been proposed: the Will and the Grace hypotheses (18).

Significance

Considering the immense economic costs associated with dishonest behavior, such as tax evasion or music piracy, reducing dishonesty is of great relevance to policy-makers. However, targeting dishonesty with interventions requires a thorough understanding of the underlying (neuro)cognitive processes. We combine neuroimaging with a task that pioneers in measuring the neural mechanisms underlying (dis)honesty. While replicating previous findings that greed drives dishonesty, we reveal that self-referential thinking processes promoted honest behavior. Moreover, we found that cognitive control does not serve the same purpose across individuals but facilitates honest decisions for cheaters, whereas it enables cheating for honest participants. We thus observe that different processes prevent dishonesty for different individuals, which can prove instrumental in the development of more effective interventions.

Author contributions: S.P.H.S., A.S., and M.A.S.B. designed research; S.P.H.S. performed research; S.P.H.S. analyzed data; and S.P.H.S., A.S., and M.A.S.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. J.D.G. is a guest editor invited by the Editorial Board.

Data deposition: Data, scripts, and images used in the task are available in Figshare (https://datarepository.eur.nl/articles/To_cheat_or_not_to_cheat_Cognitive_control_processes_override_our_moral_default/12287807).

See online for related content such as Commentaries.

[1]To whom correspondence may be addressed. Email: speer@rsm.nl.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2003480117/-/DCSupplemental.

First published August 3, 2020.

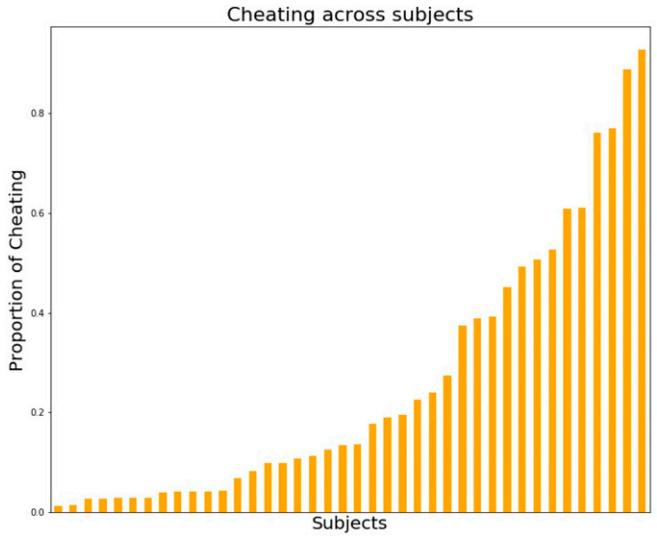

Fig. 1. Individual differences in proportion of cheating (0 to 1) on the spot-the-difference task. $n = 40$.

The Will hypothesis puts forward that people are per default selfish and dishonest and that in order to be honest, deliberate cognitive control needs to be exerted. Thus, honesty is a result of the effortful resistance of temptation, similar to the cognitive control processes that allow individuals to delay gratification (22). This hypothesis is supported by behavioral studies that have shown that participants who are cognitively depleted by demanding tasks, sleep deprived, or under time pressure are more prone to dishonest behavior (17, 20, 23, 24). Collectively, these studies suggest that people automatically serve their self-interest and require cognitive control to resist the temptation to cheat in order to maintain a positive self-image.

In contrast, the Grace hypothesis proposes that people are intuitively honest and require cognitive control to override their dominant honest impulses to occasionally profit from an opportunity to cheat. The hypothesis that cheating rather than honesty is a complex cognitive function demanding cognitive effort is supported by research showing that people react faster when asked to tell the truth as compared to lying (for meta-analyses, see refs. 25, 26) and are more honest under time pressure (27, 28). Cheating requiring cognitive capacity is also supported by findings that people cheat less when taxed by a cognitively demanding memory task as compared to a less taxing task (29). In sum, these findings suggest that honesty is intuitive and cognitive control is required to override this default intuition in order to benefit from an opportunity to cheat.

In light of these evidently contradictory findings, this study aims at investigating how cognitive control resolves the conflict between external financial rewards and one's self concept and, more specifically, how this decision process unfolds in the brain. A better understanding of the function of cognitive control in the decision to cheat may help reconcile the controversy between the Will and Grace hypotheses.

In order to study how reward, self-concept, and cognitive control influence cheating on a trial-by-trial basis, we developed an innovative task, based on a general paradigm proposed by Gai (30), in which participants could cheat repeatedly, deliberately, and voluntarily inside the MRI scanner without suspicion of the real purpose of the task. Specifically, the advantage of this task, which we call the spot-the-difference task, is that it allows for directly tracking on which trials the participants cheated, enabling us to study within subject variation in moral decisions and its neural underpinnings. Importantly, previous neuroimaging studies on cheating behavior have not been able to answer these questions as they used tasks such as the coin-flip task (5, 18), where cheating is inferred from the aggregate behavior at the end of the task, thus eliminating the possibility to study trial-by-trial variation in behavior. Notably, participants believed that the experimenter did not know that they were cheating, which is critical as participants are found to cheat less if they believe experimenters can observe their choices and know the truth

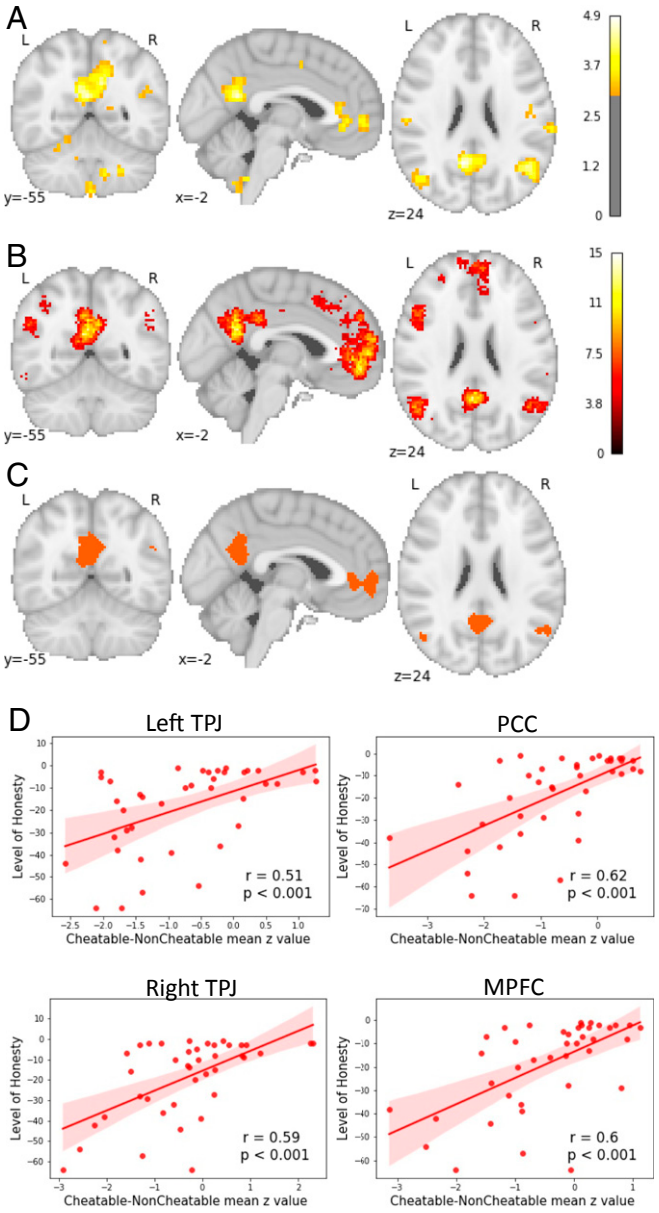

Fig. 2. Honest participants engage the self-referential thinking network more strongly than cheaters when exposed to the opportunity to cheat. (*A*) More honest participants exhibit greater activation in the PCC, MPFC, and bilateral TPJs when exposed to the opportunity to cheat. (*B*) Self-referential thinking mask obtained from Neurosynth. (*C*) Neural overlap between group level results for cheatable vs. noncheatable trials correlated with cheat count and the self-referential thinking mask obtained from Neurosynth. (*D*) The correlation between the level of honesty (reversed cheat count) and neural activation when participants were exposed to the opportunity to cheat as contrasted to no opportunity trials, for the PCC, bilateral TPJs, and the MPFC (using ROIs obtained from the conjunction analysis).

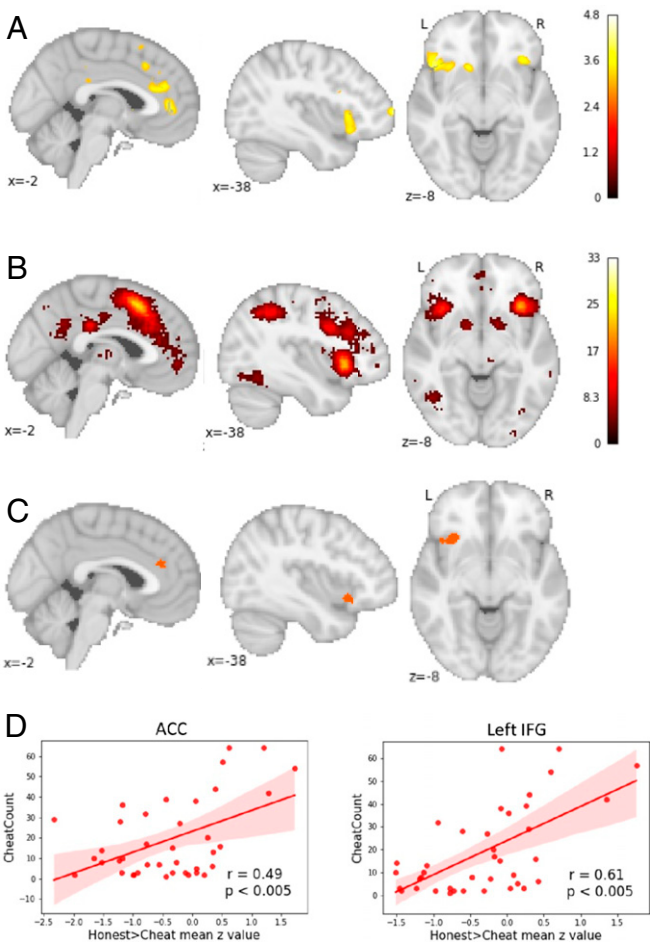

Fig. 3. Cheaters exhibit higher activation in the ACC and left IFG when deciding to be honest. (*A*) A whole-brain analysis revealed that participants who cheat more exhibit higher activation in the ACC and left IFG when deciding to be honest. (*B*) Cognitive control network derived from Neurosynth. (*C*) Neural overlap between group-level results for honest vs. cheated trials correlated with cheat count and the cognitive control mask obtained from Neurosynth in the left IFG and ACC. (*D*) The correlation between cheat count and neural activation when participants decided to be honest as contrasted to a decision to cheat, for the left IFG and the ACC (using ROIs from the conjunction analysis).

(31). The spot-the-difference paradigm is therefore a behavioral paradigm that assesses cheating behavior inconspicuously on a trial-by-trial basis enabling us to study individual differences in neurocognitive processes underlying cheating behavior while also being sensitive to within-subject variation. This will enable us to not only identify neural mechanisms promoting honesty in general but also explore the neural processes that underlie an honest decision made by someone who generally cheats.

In our analysis, we first conducted an exploratory whole-brain analysis to identify the brain networks underlying the decision to cheat or to be honest. We first identified the brain networks engaged when exposed to the opportunity to cheat and when making the decision to cheat or to be honest. To reduce the reverse inference problem (32), we then assessed the neural overlap between our results and metaanalytically derived maps associated with reward, self-concept, and cognitive control from Neurosynth (33). Subsequently, we used the regions of interest (ROIs) obtained from this conjunction analysis to conduct a trial-by-trial analysis to study the neural mechanisms underlying within-subject variation in cheating behavior and also to explore functional connectivity between the resulting networks of regions. To test the generalizability and replicability of our results, we then used cross-validation to explore whether we can use neural activation to predict unseen trials and functional connectivity patterns to distinguish between cheaters and honest participants. Here we use "cheater" and "honest" as shorthand to indicate individuals who cheated (or not) in our task, which may or may not generalize to stable traits of (dis)honesty.

We observe a central role for the cognitive control network but find that its effects depend on a person's moral default. For more honest participants, cognitive control is needed in order to cheat, whereas for cheaters, cognitive control is required to be honest. We demonstrate the generalizability and replicability of our findings by means of significant out-of-sample prediction of

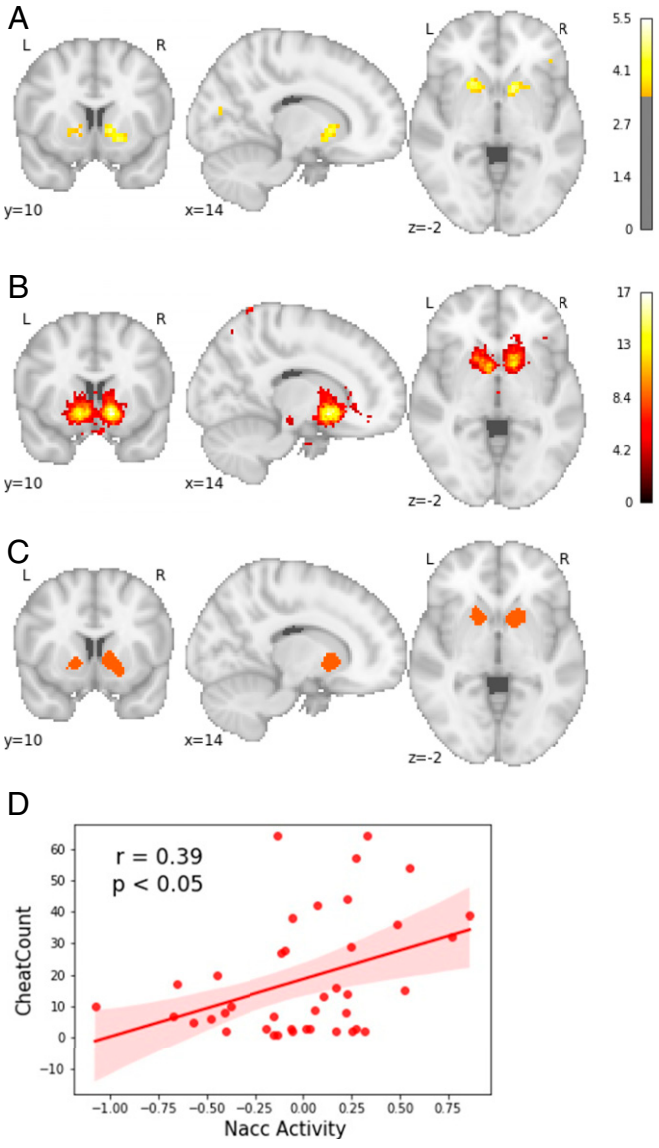

Fig. 4. Cheaters exhibit higher activity in the Nacc when making (dis)honest decisions. (*A*) The left and right nucleus accumbens are parametrically modulated by the magnitude of reward. (*B*) Reward network derived from Neurosynth. (*C*) Neural overlap between the parametric modulation analysis of the magnitude of reward and the reward anticipation network derived from Neurosynth. (*D*) Mean Nacc activity during the decision phase predicts cheat count.

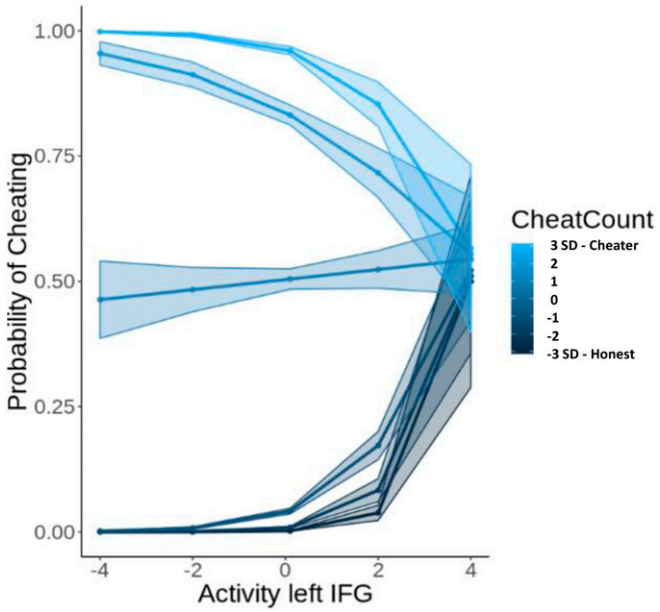

Fig. 5. Interaction effect between cheat count and the left IFG in predicting the probability of cheating. The lines that are shown are the fitted values for participants 3 SD (lightest blue), 2 SD (light blue), and 1 SD (blue) above the mean of the cheat count and participants 1 SD (dark blue), 2 SD (darker blue), and 3 SD (black) below the mean of the cheat count.

cheated decisions and cheating individuals, based on neural activation levels and connectivity patterns.

Results

Behavioral Results. Forty participants completed the spot-the-difference task (for detailed description, see *Method*) inside the MRI scanner. In the spot-the-difference task, participants were presented with pairs of images and were told that there were always three differences between the image pairs. Differences consisted of objects that were added to or removed from an image, or objects that differed in color between images. However, images could actually contain one, two, or three differences. Participants were asked to find three differences between the images. Because reward (see below) was contingent on participants reporting that they had found all three differences, without having to point them out, this design encouraged cheating behavior (i.e., reporting having found all three, even when objectively fewer than three differences were actually present in the images). Participants were told that the purpose of the study was to investigate the underlying neural mechanisms of visual search for marketing purposes such as searching for a product in an assortment or information on a webpage. Large individual differences in the total amount of cheating were observed (mean = 26%, median = 14%, SD = 26%; Fig. 1): some participants cheated only on one or two trials (17.5% of participants), whereas others only missed one or two opportunities to cheat (5%). Participants who cheated relatively often in the spot-the-difference task are from now on referred to as cheaters, whereas participants who were more honest are referred to as honest individuals. To assess suspicion about the real purpose of the study, participants were asked what the goal of the experiment was. Participants mentioned marketing research, consumer decision-making, and visual search as our general cover story suggested that visual search is important for quickly locating one's favorite brand or product in a supermarket. Importantly, none of the participants mentioned dishonesty, moral decision-making, or related concepts, which indicates that none of the participants were suspicious of the real goal of the study.

We also explored how the task characteristics of the spot-the-difference task influenced cheating behavior. Given the nested structure of our data (trials within different numbers of differences and rewards within participants), we conducted a multilevel analysis for our behavioral data. This analysis was conducted for the cheatable trials only, so all trials with three differences between the images were removed. The dependent variable was the binary response (cheating vs. honest) with a logit link (cheating = 1, honest = 0). The number of differences and level of reward served as trial level predictors. The model allowed for random intercepts and random slopes within participants. This analysis revealed a significant effect of the number of differences (excluding three differences trials) on cheating behavior ($b = 2.13$, SE = 0.44, $z = 4.85$, $P < 0.001$). This shows that participants cheated more when the crime is smaller (that is, they indicated to have found three differences more often when there were two differences as compared to when there was only one). Specifically, when there were two differences, participants cheated on 36% of the trials, whereas participants only cheated on 16% of trials with only one difference ($t = 3.28$, $P = 0.002$). No effect of reward magnitude (*Method*) on cheating behavior was observed, and no significant interaction effects between number of differences and reward were found. We also tested for possible fatigue or habituation effects by using trial number as a trial level predictor to see whether cheating behavior increased or decreased over the course of the experiment. No effects of time were observed.

Neural Mechanisms Associated with the Opportunity to Cheat. As a first step of our functional MRI (fMRI) analysis we explored the neural activation in response to the opportunity to cheat. In order to do so, we contrasted neural activity on trials in which participants had the opportunity to cheat against trials in which they did not have this opportunity (*Method* for details). To explore whether there are individual differences in the neural response to this opportunity, participants' cheat count was added as a group level covariate. The whole-brain analysis revealed that more honest participants (compared to those who cheated more) exhibited greater activation in the PCC, the MPFC, and the bilateral TPJ when exposed to the opportunity to cheat ($p_{FDR} < 0.05$; see Fig. 2*A* and *SI Appendix*, 4 for table with clusters).

As the activated network in our group-level results highly resembled the self-referential thinking network, we conducted a conjunction analysis with a metaanalytically derived self-referential thinking mask obtained from Neurosynth, false discovery rate (FDR) corrected for multiple comparisons at $P < 0.01$ (33) (Fig. 2*B* and *SI Appendix*, 3) to test whether there is indeed neural overlap. Neural overlap was found in the PCC (overlap [mm^3] = 4,600), in the MPFC (overlap [mm^3] = 4,072), in the right TPJ (overlap [mm^3] = 869), and in the left TPJ (overlap [mm^3] = 608) (Fig. 2*C*).

Table 1. Multilevel logistic regression model using the cognitive control network to predict cheating

	Estimate	SE	z value	Pr(>|z|)
Intercept	−1.582	0.069	−22.793	<0.001
ACC	0.132	0.06	2.306	0.02
L IFG	0.422	0.061	6.908	<0.001
Cheat count	1.60	0.07	23.735	<0.001
L IFG × cheat count	−0.382	0.062	−6.192	<0.001

The source of anatomical labels: Automated Anatomical Labeling Atlas tools cross referenced with Neurosynth. L, left.

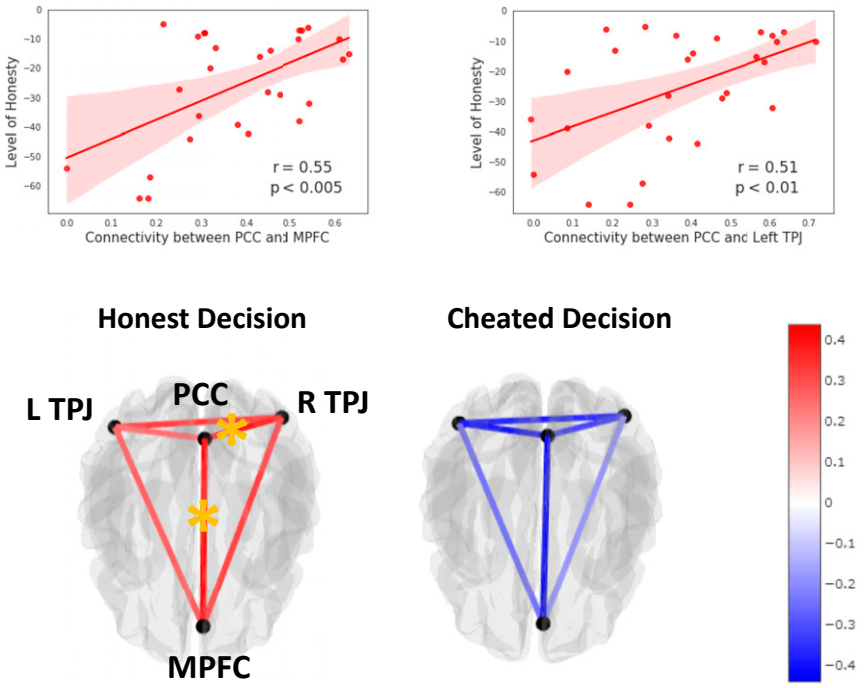

Fig. 6. (*Top*) Correlation between level of honesty (reverse cheat count) and functional connectivity between PCC and MPFC and PCC and left TPJ. (*Bottom*) Connectome showing the correlation between level of honesty and the functional connectivity within the self-referential thinking network during (*Left*) honest decisions and (*Right*) cheated decisions. The red lines represent a positive correlation between regions, whereas the blue lines represent a negative correlation. Darker color represents stronger correlation. Orange stars indicate a significant correlation after correction for multiple comparisons (FDR at $P < 0.05$).

Neural Mechanisms Underlying the Decision to Cheat. Next, we explored which neural mechanisms underlie the decision to cheat or not, when given the opportunity. To answer this question, we contrasted the neural activation of trials where participants had the opportunity to cheat but decided to be honest, against trials on which participants decided to cheat. As before, to explore whether there are individual differences in the neural processes underlying honest as compared to dishonest decisions, participants' cheat count was added as a group-level covariate.

We found that participants who cheated more showed higher activity in the anterior cingulate cortex (ACC) and the inferior frontal gyrus (IFG) when they made the decision to be honest ($P < 0.001$, uncorrected; see Fig. 3*A* and *SI Appendix*, 5 for table with clusters). Stated differently, cheaters engage their ACC and IFG more than honest participants when refraining from cheating. As the activated network in our group-level results highly resembled regions within the cognitive control network, we conducted a conjunction analysis with a metaanalytically derived cognitive control mask obtained from Neurosynth (33) (Fig. 3*B* and *SI Appendix*, 3) to test whether there is indeed neural overlap. Neural overlap was found in the ACC (overlap [mm^3] = 168) and in the left IFG (overlap [mm^3] = 1,256) (Fig. 3*C*).

Neural Correlates of the Sensitivity to Reward Are Associated with Cheating.

Level of difficulty phase. Although we did not find any effects of reward on cheating on the behavioral level, we did want to test whether the participants responded to the reward on the neural level, as previous research has eluded to the relevance of reward anticipation in explaining individual differences in cheating (4, 5). Here we investigated whether participants were motivated by

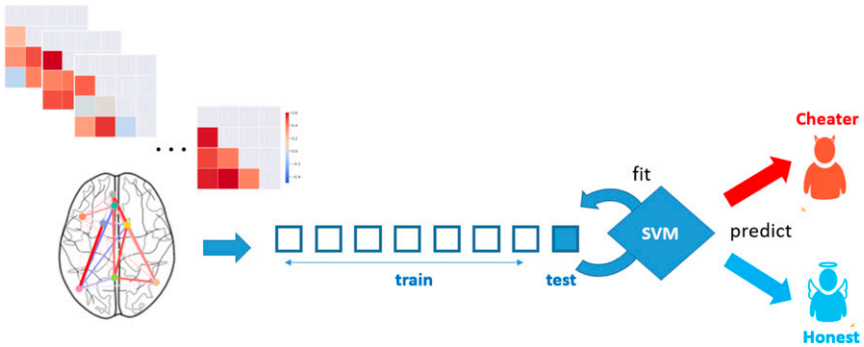

Fig. 7. Using participants' connectivity patterns within the self-referential thinking network during decision-making to classify participants as cheaters or honest participants using support vector classifiers implemented with eightfold cross-validation.

the possible rewards that could be obtained on each trial and whether participants differentiated between the different magnitudes, 5, 20, and 40 cents, of reward on the neural level. We conducted a parametric modulation analysis where we used the onsets of the level of difficulty phase of each trial and added the magnitude of reward at each trial as a parametric modulator on the first level. The analysis revealed that the magnitude of reward modulated the activity in the bilateral Nacc significantly ($P_{FDR} < 0.05$; see Fig. 4*A* and *SI Appendix*, 6 for table with clusters).

As the Nacc is well known for its role in processing the anticipation of reward (34–36) this suggests that the participants were indeed motivated by the potential rewards presented at the beginning of the trial. Further, differences between levels of reward magnitude were reflected in different levels of activity in the Nacc, suggesting that participants were indeed differentiating between the different reward magnitudes.

As the activated network in our second-level results highly resembled the reward anticipation network, we conducted a conjunction analysis with a metaanalytically derived reward anticipation mask obtained from Neurosynth with FDR corrected for multiple comparisons at $P < 0.01$ (33) (Fig. 4*B* and *SI Appendix*, 3) to test whether there is indeed neural overlap. Neural overlap was found in the right Nacc (overlap [mm^3] = 2,040) and left Nacc (overlap [mm^3] = 840) (Fig. 5*C*). We also conducted an additional second-level analysis, in which we added the cheat count as a covariate, in order to explore whether reward sensitivity in the level of difficulty phase differed between subjects. However, no significant differences were observed, indicating that participants were equally sensitive to the rewards, independent of how often they cheated.

Decision phase. To explore how the effect of reward anticipation, as represented by activity in the Nacc, on cheating differs for cheaters and more honest participants, we then used the ROIs derived from the conjunction analysis between our parametric modulation analysis and the Neurosynth map for reward (Fig. 4*C*) and regressed mean Nacc activity per subject during the anticipation and decision phase against the cheat count. This analysis revealed that average Nacc activity significantly predicted cheat count ($b = 18.29$, SE = 7.01, $P < 0.05$; Fig. 4*D*) during the decision phase, whereas no significant effect was found during the level of difficulty phase ($b = -8.89$, SE = 14.2, $P = 0.54$). This suggests that participants are equally sensitive to reward during the level of difficulty phase when there is no moral conflict; however, when making the decision to cheat (or be honest), participants who cheat more seem to be driven more strongly by anticipation of reward.

Investigating within-Subject Variation in Cheating: Trial-by Trial Analysis. In order to further explore how self-concept, reward, and cognitive control influence decisions to cheat, we conducted a trial-by-trial analysis, which allowed us to investigate the neural mechanisms that determine why the same person may cheat on some occasions and remain honest on others. As a first step, we extracted average trial-by-trial activation from individual regions within the reward, cognitive control, and self-referential thinking network, using the conjunction between our second-level results and the Neurosynth maps (Figs. 2*C*, 3*C*, and 4*C*), resulting in one data matrix where the rows represent trials and the columns represent the regions of interest. Given the nested structure of our data (trials within different number of differences and rewards within participants) we then conducted a multilevel analysis for each of the networks (self-referential thinking, cognitive control, and reward). The dependent variable was the binary response with a logit link (cheating = 1, honest = 0). The averaged activity within the obtained regions of interest served as trial-level predictors, whereas the cheat count served as a subject-level predictor. The models allowed for random intercepts and random slopes within participants.

Assessing the relative importance of the networks. To investigate which of the networks is most important in predicting cheating on the trial level, we performed variable selection for generalized linear mixed models by means of L1-penalized estimation. This was implemented using the glmmlasso package in R, which implements a gradient ascent that allows us to maximize the penalized log-likelihood, yielding models with reduced complexity (37). The lasso regression adds a penalty term to the equation which shrinks less important coefficients in the model to zero and thus reduces complexity of the model and multicollinearity of predictors (38). In this way it also selects the most important predictors in the model. This analysis revealed that the ACC ($b = 0.13$, SE = 0.06, $P = 0.02$), the left IFG ($b = 0.42$, SE = 0.06, $P < 0.001$), the cheat count ($b = 1.59$, SE = 0.07, $P < 0.001$), and the interaction effect between the left IFG and the cheat count ($b = -0.38$, SE = 0.06, $P < 0.001$) were most important in predicting cheating (Table 1). These results suggest that the cognitive control network is most important in predicting cheating on the trial level. Inspecting the plot of the interaction effect (Fig. 5), we see that for participants who cheat a lot (light blue lines), higher levels in the left IFG are associated with lower probabilities of cheating, whereas for more honest participants (dark blue lines), higher activity in the left IFG is associated with higher probability of cheating. These findings suggest that the effect of the left IFG on cheating depends on whether a participant has the general tendency to cheat or to be honest.

Testing the predictive accuracy of the model. As the cognitive control regions were found to be most predictive of cheating, we used these predictors to test the prediction accuracy of our model. In order to do this, we used the trial-level activation in the ACC and left IFG, excluding the cheat count, obtained from the conjunction analysis and trained a multilevel logistic regression model, with random slopes and intercepts, on a training set (70% of the data). Subsequently, we tested the model on the left-out 30% of the data. As the dependent variable, cheating, was imbalanced, we used two accuracy metrics that are insensitive to the class imbalance, namely, the area under the curve (AUC) and the F1 score, which is the harmonic mean of the precision and recall. Statistical significance was estimated using permutation tests where the dependent variable (cheating) was permuted 5,000 times and the classification metrics were estimated based on random permutations. We found that we were able to significantly predict cheating based on unseen data from activity in the cognitive control network (AUC = 76%, F1 = 89%, $P < 0.001$).

Individual Differences in Functional Connectivity during Decision-Making.

Connectivity within the self-referential thinking network. In order to further explore how the different areas resulting from the different contrasts described above interact with each other during decisions to cheat, we investigated the functional connectivity between these areas during the decision phase of the

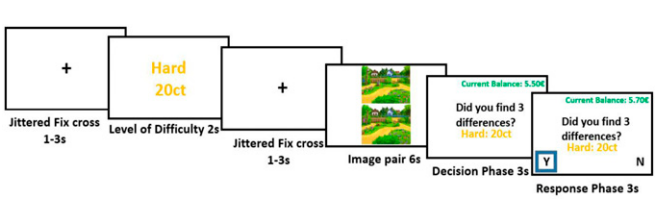

Fig. 8. One trial of the spot-the-differences paradigm. Participants view a screen indicating the difficulty and value of the trial, then the image pair appears for 6 s, and then participants have to indicate whether or not they spotted all three differences.

spot-the-difference task. To avoid the problem of activation-induced correlations we implemented beta-series correlations (39) (see *Method* for details). The beta-series correlation analysis revealed that functional connectivity within nodes of the self-referential thinking network were more strongly connected for honest participants than for cheaters when making honest decisions. Specifically, correlations between honesty and functional connectivity were found between the PCC and left TPJ ($r = 0.51$, $p_{adj} < 0.05$) and between PCC and MPFC ($r = 0.55$, $p_{adj} < 0.05$; Fig. 6). No significant correlations between honesty and functional connectivity were found for cheated decisions. In addition, the correlation between honesty and functional connections between PCC and left TPJ and between PCC and MPFC during honest decisions were significantly different from the correlation during cheated decisions (both comparisons $z > 2$, $p_{adj} < 0.005$). Thus, the nodes within the self-referential thinking network, particularly between MPFC, left TPJ, and PCC, seem to be more intimately connected to promote honesty particularly for honest participants, whereas when the connectivity between these nodes breaks down, honest participants tend to cheat.

Classification of cheaters vs. honest participants based on functional connectivity patterns. To test whether there is sufficient information in the connectivity patterns within the self-referential thinking network reported above to predict individual differences in honesty, a support vector classifier (40, 41) with linear kernel ($C = 1$) was trained on the functional connectivity patterns of each participant to determine whether a participant was a cheater or an honest participant (categorized by median split). In order to avoid overfitting and inflated prediction accuracy (42) this was done using eightfold cross validation (Fig. 7). Significance was estimated using permutation testing ($n = 5{,}000$). The classification analysis revealed that we could significantly classify an unseen participant as a cheater or an honest individual based on the connectivity patterns within the self-referential thinking network (accuracy = 71%, F1 = 75%, $P < 0.05$).

Discussion

In this study we explored how neural mechanisms associated with reward anticipation, self-referential thinking, and cognitive control determine the (dis)honesty of individual decisions. Using the spot-the-difference task to study trial-by-trial cheating behavior we found that the effect of cognitive control depends on a participants' inclination to be honest or dishonest, in other words, on their moral default.

We found that more honest participants engaged a network of brain regions associated with self-referential thinking when exposed to the opportunity to cheat. Particularly, participants who were generally honest exhibited higher activity in the self-referential thinking network composed of the PCC, the bilateral TPJs, and the MPFC. We provided further evidence that our results indeed reflect self-referential thinking processes by means of a conjunction analysis with Neurosynth data. Exploring the functional connectivity within the self-referential thinking network, we also found that more honest participants exhibited stronger connectivity during honest decisions between all nodes in this network, whereas this connectivity within the self-referential network broke down during cheated decisions. Collectively, these findings highlight the importance of our moral self-concept and related self-referential thinking processes in promoting honesty.

In line with previous research (4, 5), we found that cheaters exhibited stronger sensitivity to reward during decision-making. Our results revealed that all participants were anticipating reward and were sensitive to differences in magnitude of reward during the initial phase of the trial, where the potential reward for finding the differences between the two images is presented, without any presence of moral conflict. However, cheaters, as compared to more honest participants, were more strongly driven by reward when making the decision whether to cheat or not. Specifically, cheaters exhibited higher neural activation in the Nacc, which is an area that has been consistently linked to reward anticipation (34–36), during the decision phase. Thus, whereas all participants are sensitive to differences in the magnitude of reward in the absence of moral conflict, particularly the cheaters are driven by the anticipation of reward when making the decisions to cheat.

Importantly, our study suggests that the function of cognitive control depends on a person's moral default. Particularly, we found that for honest participants, more cognitive control, as represented by higher activity in the left IFG, was needed to cheat, whereas for participants who cheated frequently, control was needed in order to be honest. While honest participants needed cognitive control to overcome their inclination of being honest in order to cheat, cheaters had to exert control to override their greedy tendencies in order to be honest. Thus, our analyses indicated that the role of cognitive control depends on a person's moral default.

In the literature, there has been a debate between proponents of the Will hypothesis and the Grace hypothesis. Research supporting the Will hypothesis (17, 20, 43) suggests cognitive control is needed to be honest. In direct opposition to this, another stream of research has accumulated evidence in favor of the Grace hypothesis (for metaanalyses, see refs. 18, 24–27), advocating that cognitive control is required for dishonesty.

Our findings help reconcile this conflict as they suggest that people are distributed along a continuum, from individuals who are generally honest to participants who can be considered cheaters. Participants on one side of the spectrum have a default inclination to be honest which is associated with more self-referential thinking when given the opportunity to cheat. In contrast, individuals on the other side of the spectrum have a default inclination for dishonesty, and their decisions seem to be driven more strongly by rewards. In order to achieve and maintain a subjectively justifiable balance where one can occasionally profit from cheating but still maintain a positive self-image, people on both sides of the spectrum sometimes need to overcome their initial impulse and default behavior. A generally honest person will need to overcome the default of being honest in order to profit from cheating from time to time, whereas a cheater needs to inhibit the predominant selfish response in order to occasionally be honest and maintain their self-concept.

Thus, it appears that the effect of cognitive control depends on our moral default. For honest people the Grace hypothesis applies: honesty results from the absence of temptation, and response inhibition is needed to cheat. In contrast, for cheaters the predictions of the Will hypothesis apply, and active resistance of temptation in the form of inhibition is needed to be honest. Extending findings from cognitive psychology to the social/moral domain, our results suggest that cognitive control seems to serve the purpose of overriding our default behavior. We show that cognitive control processes, well established through previous research on response inhibition (44–46), can serve different purposes for different people in the context of (dis)honest decision-making. Our study thus contributes to the reconciliation of the controversy on the role of cognitive control in moral decision-making.

In addition, our findings also point to the importance of self-referential thinking processes and the maintenance of a positive self-concept. Whereas previous neuroimaging research has mainly focused on the role of cognitive control and reward sensitivity in cheating behavior, our study finds neural evidence in favor of the self-concept maintenance theory (13). Our results indicate that besides reward and control processes, self-referential thinking as represented by activation in the PCC, MPFC, and bilateral TPJs was engaged, particularly in honest

participants, when they were tempted to cheat and more strongly functionally connected when making honest decisions. Thus, our neural evidence suggests that when exposed to an opportunity to cheat, particularly honest people do value their moral self-concept and its maintenance enough to forgo potential financial gains.

Our interpretation of the activation in the PCC, MPFC, and bilateral TPJs in terms of self-referential thinking was supported by the observed overlap with a metaanalytically derived activation map. In addition, it has been found that functional connectivity between these areas (14, 15) during rest is associated with higher-level metacognitive operations such as self-reflection and introspection (47). Moreover, resting-state functional connectivity between these areas correlates positively with ratings of internal awareness (48). Based on these findings, we can be fairly confident that the observed network of brain areas can indeed be viewed as neural correlates of the self and self-referential thinking (49).

Nonetheless, it has to be noted that the same network of regions has also frequently been associated with other psychological processes, most notably theory of mind (50–52) and mental time travel (53, 54). While these two processes are distinct from self-referential thinking, they may nonetheless share the common underlying mechanism of perspective taking, as they all require projection of one's self to an alternative perspective of time or person (55). In our study it seems unlikely that the activity in this network of regions reflects theory-of-mind processes as our task was designed so that there is no identifiable victim of the dishonest decisions that can be empathized with. Similarly, it seems less plausible that (particularly honest) participants were recollecting past experiences or made plans for the future while making honest decisions. Therefore, we surmise that in our experiment, the observed areas reflect perspective-taking that is required for self-referential thinking.

To examine the generalizability of our findings, we also tested the predictive power of the cognitive control regions in predicting cheating on a trial-by-trial basis using cross-validation. We found that we could significantly predict with high accuracy on unseen data whether on a given trial participants would be honest or would cheat. Moreover, to assess whether connectivity patterns between the different networks contained relevant information about individual differences in honesty, we used support vector classifiers trained on participants' connectivity patterns to discriminate cheaters from honest participants and found that we could indeed accurately classify whether a participant is a cheater or not. Combining the two models did not significantly increase trial-by-trial prediction of cheating (*SI Appendix*, 9). From the perspective of scientific rigor, cross-validation is a more conservative way to infer the presence of a brain–behavior relationship as compared to correlation or regression, as it is designed to protect against overfitting by testing the strength of the association in a new sample. This increases the probability of successful replication in future studies.

Nevertheless, participants who cheated or were honest on our task may not act similarly in a different context. Their (dis) honesty, as measured in our task, may thus not reflect a stable moral trait. A recent study using resting state fMRI, however, seems to support the notion that the observed individual differences in honesty may generalize (56). That study shows that (dis)honesty, as observed in our task, is represented in stable functional connections at rest, within the self-referential thinking network and between the self-referential thinking and the reward and cognitive control network. Resting-state functional connectivity has been shown to be robust and reliable over time and across tasks, which may suggest that the observed correlation with cheating behavior may generalize as well. Further, these individual differences in cheating were also found to be strongly correlated with stable personality characteristics related to impulsivity.

In order to rule out alternative explanations for our findings, we conducted several control analyses. First, in order to test whether neural differences during the decision phase were not driven by differences in levels of engagement with the task, we explored the neural processes during the visual search phase of each trial. As expected for a visual search task, we found that participants showed increased activation in areas related to visual and cognitive processing, working memory, and navigation while searching for the differences (*SI Appendix*, 7). Importantly, no significant differences in neural activations during visual search were found between honest participants and cheaters. This eliminates the possibility that our neural findings were confounded by processes related to differences in engagement or effort during visual search. Second, we also conducted an exploratory factor analysis, which revealed that regions of interest used in our trial-by-trial and functional connectivity analyses indeed belonged to three separate networks that could be clearly identified as the control, reward, and self-referential thinking network (*SI Appendix*, 8).

In reference to previous neuroimaging research on moral decision-making, our findings align with the early work using hypothetical moral dilemmas (57), instructed lying paradigms (58, 59), and work using the die-roll task (18) in highlighting the importance of the cognitive control network, including areas such as the ACC and IFG, in moral decision-making. As stated above, our findings are also in line with those of Abe and Greene (5), converging on the conclusion that a more sensitive and responsive reward network is associated with higher levels of cheating.

It is worth noting, however, that reward does not always increase cheating. While higher payoffs were found to increase cheating in sender–receiver games and to some extent in the coin-flip task, they did not have an effect on the die-roll or matrix task (for review, see ref. 60). An explanation for this inconsistency may be that reward size may have different, perhaps opposing effects for different individuals (61). This aligns with our findings, suggesting that while for cheaters, higher rewards mainly increase the temptation to cheat, for honest participants they may increase the moral cost and associated guilt of cheating, due to the increase in magnitude of the transgression (13, 60).

With regard to neural processes linked to self-referential thinking, an fMRI study by Greene et al. (62) found that a network of regions including the MPFC, PCC, and bilateral TPJ were involved in making judgements about more personal as opposed to abstract hypothetical moral dilemmas, which they attributed to general emotional processes. More recently, a metaanalysis on neuroimaging research on moral decision-making conducted by Lisofsky et al. (63) reported that experimental deception paradigms that involved an identifiable victim and consequently perspective taking were associated with increased activation in the right temporal parietal junction and the bilateral temporal pole, which have been associated consistently with theory of mind processes (64) as compared to less interactive deception and cheating studies. Based on these findings, Lisofsky et al. (63) argue that, particularly in studies involving social interaction and an identifiable victim, not only control processes but also perspective-taking and moral reasoning processes are important. Our findings add to their conclusion by demonstrating that also in contexts without an identifiable victim, a similar network of regions, involving the TPJ but also the MPFC and PCC, is crucial in determining the outcome of moral decisions. This suggests that similar neural mechanisms may underlie self-referential thinking and perspective-taking processes in the context of moral decision-making.

To conclude, we used a task that allows measuring cheating on the trial level in an fMRI environment. Using this task, we found

that not only reward sensitivity but also the extent to which someone engages self-referential thinking processes determines whether someone is a cheater or tends to be honest most of the time. Importantly, we also found that the role of cognitive control in (dis)honesty depends on a person's moral default. These findings may prove to be useful for developing interventions targeted at reducing cheating and dishonesty. Considering the huge economic costs caused by dishonest behavior, such as tax evasion, music piracy, or business scandals such as the Volkswagen emission fabrications, reducing dishonest behavior effectively is of great relevance to policy-makers and our economy in general.

Taken together, we showed that the neural mechanisms engaged in (dis)honest decisions, ranging from neural activation in reward, self-referential thinking, and control networks to functional connectivity patterns, differ fundamentally between honest and dishonest participants. Specifically, we found that cognitive control overrides a person's moral default. Cognitive control allows honest people to cheat at times, whereas it enables cheaters to sometimes be honest. These insights contribute to a deeper understanding of the neural correlates of individual differences in moral decision-making. Future research may explore whether neural markers associated with dishonesty are also observable in more stable neural measures such as resting state functional connectivity or structural brain differences.

Method

Participants. The reported analyses are based on 40 participants (30 females; age 18 to 35 y; $M = 23.7$, SD = 3.2) recruited from an online community for university students, where students can sign up for experiments. An initial screening interview ensured that all participants were right-handed with normal or corrected to normal vision, spoke English fluently, were not on any psychoactive medication influencing cognitive function, and had no record of neurological or psychiatric illness. The study was approved by the Erasmus Research Institute of Management (ERIM) internal review board and was conducted according to the Declaration of Helsinki.

Task and Stimuli.

Spot-the-difference task. In the spot-the-difference task, participants were presented with pairs of images and were told that there were always three differences between the image pairs. Differences consisted of objects that were added to or removed from an image or objects that differed in color between images. However, images could actually contain one, two, or three differences. Participants were asked to find three differences between the images. Because reward (see below) was contingent on participants reporting that they had found all three differences, without having to point them out, this design encouraged cheating behavior (i.e., reporting having found all three, even when objectively fewer than three differences were present in the images).

Participants were told that the purpose of the study was to investigate the underlying neural mechanisms of visual search for marketing purposes such as searching for a product in an assortment or information on a webpage. In order to increase credibility of this cover story a simple visual search task was added at the beginning of the experiment (*SI Appendix*, 1), which was also performed in the scanner while participants were undergoing localizer scans. Further, participants were instructed that the neurocognitive effect of motivation, elicited by monetary reward, on speed and accuracy of visual search was investigated. Although participants were told that there were three differences in all trials, in 25% of the trials, there were only two differences, and in 25%, there was only one difference. All stimuli were standardized in size and were presented on a white background on a computer screen. The ratio of 50 to 50% (three differences vs. fewer than three differences) was chosen based on the results of pilot studies that indicated this ratio to be optimal in reducing suspicion that the pairs did not always contain three differences.

Trials were further categorized into normal (50%), hard (25%), and very hard trials (25%), for which participants could receive 5, 20, and 40 cents, respectively. All of the trials with three differences (the filler trials) were categorized as normal trials, whereas trials with fewer than three differences (the trials of interest) were randomly categorized as hard or very hard trials. Consequently, the reward was independent of the number of differences in the image pair for the trials of interest, which is important in order to be able to disentangle the effects of reward and cheating magnitude (the actual number of differences) on cheating behavior. The different levels of difficulty were added to reduce suspicion about the real purpose of the task. It was assumed that if trials are labeled as hard or very hard, it would be more credible to the participant that the image pair actually contained three differences, but they were just too hard to spot. In addition, levels of difficulty were introduced to eliminate possible demand effects: we wanted participants to cheat for monetary reward and not to prevent seeming incompetent, which may be associated with different underlying neural mechanisms and consequently confound the analysis.

To further reduce suspicion about the purpose of the study, ~10% of all trials were point-and-click trials. In these trials, participants had to click on the location in the images where they spotted the differences using a joystick. As a consequence, cheating was not possible on the point-and-click trials. Participants always knew prior to the start of a trial whether it was a point-and-click trial indicated by a screen requesting participants to click on the image. This ensured that participants would not refrain from cheating on all other trials, while still reducing the suspicion about the real purpose of the study. Participants were told that only 10% of trials were point-and-click trials because it would take too much time to point out the differences for every pair. Further, participants were instructed that excessive movement by manipulating the joystick would interfere with the brain signal. In sum, there were 144 regular trials (of which 72 cheatable trials) and 12 point-and-click trials. The maximum amount of money earned, in case a participant cheated on all cheatable trials, was ~ €35, whereas in case a participant would not cheat at all, he or she would earn ~ €7.50. After completion of the full study, participants were debriefed that the purpose of the study was to investigate the underlying neural mechanisms of (dis) honest decision-making. They were informed that the number of differences between pictures and level of reward were manipulated to encourage cheating. To be fair to all participants, they were all paid out the maximum amount, irrespective of their actual cheating behavior. In addition, participants received a flat fee of €10 for participation in the scanning session.

Each trial started with a fixation cross which was presented for a variable amount of time between 1 and 3 s (Fig. 8). Subsequently, the level of difficulty screen was presented for 2 s informing the participants about the level of difficulty of the upcoming trial. This screen also displayed how much money could be earned on that trial. As a result, participants were constantly aware of the potential gains of cheating. Next, an image pair was presented for 6 s, a length determined by the behavioral pilots, and participants engaged in the visual search. Afterward, the participants were asked whether they spotted all three differences (yes/no response). On this decision phase screen, again the potential reward for this trial was presented, in order to make the reward more salient and increase cheating behavior. After 3 s, the response phase started in which participants' responses were recorded. In the decision phase and the response phase the current balance was also shown, which was done to demonstrate to the participants that if they stated that they had found the three differences, their current balance increased immediately. It was assumed that this direct noticeable effect of behavior on the increase of the current balance would further motivate participants to cheat.

The decision phase and response phase were separated to isolate the decision from motor responses. This was important for the fMRI analysis as we wanted to isolate the neural mechanisms underlying decision-making from possible neural confounds related to button presses. Besides that, the buttons corresponding to "yes" and "no" were switched across trials to further reduce confounding effects and to reduce the response bias for the dominant hand. Once the participants responded, the choice was highlighted by a blue box for 500 ms to indicate that the response was recorded, and the trial ended. If no response was made, the trial ended after 3 s. In addition, there were five practice trials, in which participants could get acquainted with the task. Stimulus presentation and behavioral data acquisition was performed using Presentation software (Version 18.0, Neurobehavioral Systems, Inc., www.neurobs.com).

The main advantage of our experimental design is that it allowed tracking on which trials the participants cheated. As we knew how many differences there were in each image pair, we knew precisely whether the participants cheated or not. Further, by varying the number of differences, this design enabled us to assess the magnitude of cheating (i.e., cheating when only one vs. two differences were found). It is therefore a behavioral paradigm that allows us to assess cheating behavior inconspicuously on a trial-by-trial basis in the scanner.

Stimuli. Stimuli for the task consisted of 144 spot-the-difference image pairs that were downloaded from the Internet. Cartoon images of landscapes

containing several objects were selected, to make them engaging and challenging enough for the participants. Landscapes were chosen as they generally satisfied the necessary criterion of containing several different objects. The stimuli consist of pairs of images that are identical apart from a certain number (one to three) of differences that were created using Adobe Photoshop. Differences consisted of objects added to or removed from the landscape picture or changed colors of objects. Differences were fully randomized across all pairs of images, which means that all image pairs could be presented with either one, two, or three differences. To make sure that participants would be able to find the differences between the images in a reasonable amount of time and to minimize the chance of participants believing that they had seen a difference when they had not (false positives), we ran a pilot study on Amazon's Mechanical Turk (n = 205) to test the difficulty to spot the differences between the images and to determine the optimal duration of picture presentation (*SI Appendix, 2*).

Experimental procedure. Before the experiment started, participants were introduced to the cover story, the tasks, and the scanner environment, and they signed the informed consent form. They were then informed about and checked on the safety requirements for MRI scanning and completed practice trials for both visual search tasks outside of the scanner. Subsequently, they were guided into the scanner and completed the simple visual search task (5 min) followed by the spot-the-difference task which took ∼45 min. Scans for the spot-the-difference task were acquired in six separate runs to allow participants short periods of rest in order to reduce head movements during the scans and to ensure that participants were attentive throughout the task. After completing the two tasks in the scanner, participants were taken to a separate room in absence of the experimenter and filled out a short questionnaire including questions about their thoughts on the purpose of the task.

FMRI Acquisition. The fMRI images were collected using a 3T Siemens Verio MRI system. Functional scans were acquired by a T2*-weighted gradient-echo, echo-planar pulse sequence in descending interleaved order (3.0 mm slice thickness, 3.0 × 3.0 mm in-plane resolution, 64 × 64 voxels per slice, flip angle = 75°). TE was 30 ms, and TR was 2,030 ms. A T1-weighted image was acquired for anatomical reference (1.0 × 0.5 × 0.5 mm resolution, 192 sagittal slices, flip angle = 9°, TE = 2.26 ms, TR = 1,900 ms).

fMRI Analysis.

Preprocessing. The fMRI data were preprocessed using fMRIPrep version 1.0.8, a Nipype based tool (65). The reason for choosing fMRIPrep was that it addresses the challenge of robust and reproducible preprocessing as it automatically adapts a best-in-breed workflow to virtually any dataset, enabling high-quality preprocessing without the need of manual intervention (66). Each T1w volume was corrected for intensity nonuniformity and skull-stripped. Spatial normalization to the International Consortium for Brain Mapping 152 Nonlinear Asymmetrical template version 2009c (67) was performed through nonlinear registration, using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white matter (WM), and gray matter was performed on the brain-extracted T1w. Field map distortion correction was performed by coregistering the functional image to the same-subject T1w image with intensity inverted (68) constrained with an average field map template (69). This was followed by coregistration to the corresponding T1w using boundary-based registration (70) with 9 degrees of freedom. Motion-correcting transformations, field distortion correcting warp, blood-oxygen-level–dependent images-to-T1w transformation, and T1w to template Montreal Imaging Institute (MNI) warp were concatenated and applied in a single step using Lanczos interpolation. Physiological noise regressors were extracted applying CompCor (71).

Principal components were estimated for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). Six tCompCor components were then calculated including only the top 5% variable voxels within that subcortical mask. For aCompCor, six components were calculated within the intersection of the subcortical mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run. Frame-wise displacement (72) was calculated for each functional run using the implementation of Nipype. For more details of the pipeline, see https://fmriprep.org/en/latest/workflows.html.

Statistical analyses. For each participant we estimated a general linear model (GLM) using regressors for onsets of the decision phase for cheated trials, honest trials, cheatable trials (trials with fewer than three differences), and noncheatable trials (trials with three differences). The duration of the epoch for the decision phase was 3 s, and the beginning of the decision phase was used as onset time. The decision phase was used as it provides all of the necessary information to make the decision and is free of brain activity related to motor responses. In addition, regressors were added for the onsets of the level of difficulty phase with a separate regressor for each level of reward. For the level of difficulty phase the duration was 2 s. This phase was used to test whether participants are indeed sensitive to differences in potential gains, as it provided information about the possible reward without any moral conflict. Besides that, in order to ensure that there were no significant differences in engagement or motivation in the spot-the-difference task between conditions or subjects, regressors were added for the onsets of the visual search phase in which the image pairs were presented on the screen. The duration of the visual search phase was 6 s (Fig. 1). Last, regressors for the button presses were added. Average background, WM and CSF signal, framewise displacement, six head motion regressors, and six aCompCor regressors, all obtained from fMRIprep, were entered as regressors of no interest. All regressors were convolved with the canonical hemodynamic response function. A smoothing kernel of 5 mm full width at half maximum was applied. Linear contrasts were computed between honest and cheating decisions and between cheatable and noncheatable trials. These contrasts were then subjected to a random effects analysis to compute main effects (one sample t test) and to regression analyses with behavioral data (i.e., total amount of cheating for each participant) as regressors.

Cheatable vs. noncheatable trials. To identify the neural correlates associated with the opportunity to cheat, we contrasted the neural activation during cheatable trials (trials with fewer than three differences), against activation in noncheatable trials (trials with three differences) in both directions. Subsequently, using the contrast images obtained for each subject, one-sample t tests were conducted on the group level to explore the average effect of being exposed to the opportunity to cheat across participants. We also added the cheat count, which is a measure how often each participant cheated in total on the spot-the-difference task, as a group-level covariate to explore whether there are individual differences in the neural mechanisms when exposed to the opportunity to cheat, between individuals who cheat a lot vs. those who rarely cheat. The threshold applied to the group level statistical maps was a voxel-wise FDR of $P < 0.05$ to correct for multiple comparisons. Clusters of activation resulting from the thresholding were characterized in terms of their peak voxels in the MNI coordinate space.

Honest decisions vs. cheating. To explore the neural mechanisms underlying the decision to cheat, we contrasted neural activation in the decision phase on trials on which participants cheated against trials in which they did not, in both directions. For each of these contrasts we then conducted one-sample t tests on the group level to explore the average effects of each of these contrasts across participants. In addition, we also entered the total cheat count for each participant as covariate on the group level to investigate the correlation between behavior and neural activation in the contrasts of interest. Based on the resulting beta images, second-level random-effects group contrast maps were then created in both directions (i.e., positive and negative correlation between activation and cheat count). The threshold applied to the group-level statistical maps was a voxel-wise FDR of $P <$ 0.05 to correct for multiple comparisons. Clusters of activation resulting from the thresholding were characterized in terms of their peak voxels in the MNI coordinate space. Due to the fact that participants engaged in spontaneous, voluntary, and deliberate cheating, the proportion of cheated and honest trials was not balanced for most of the participants. To account for possible confounding statistical effects of this imbalance, we undersampled the majority class for each participant to create a perfect balance when estimating the contrasts (73).

Single-trial activation estimation. An important contribution of our task is that it allows us to assess cheating behavior on a trial-by-trial basis. That is, we are able to assess why a person who is generally honest decides to cheat on some trials and why a cheater might refrain from cheating on some occasions. To explore which neural mechanisms underlie this within-subject variability, we extracted the neural activation from the ROIs identified in the analyses described above during decision-making for each trial for each subject. These trial-by-trial activations could then be fed into multilevel models to explore which neural mechanisms may explain within subject variability.

To obtain single-trial neural activations for the trial-by-trial multilevel models, individual time series were modeled using a double γ hemodynamic response function in a single-trial GLM design using FMRIB Software Library's fMRI Expert Analysis Tool. Specifically, one GLM fitted a hemodynamic response function for each trial, following the least-squares all approach (74), using the decision phase and level of difficulty phase of each trial, resulting in parameter estimates of sustained activations for each trial for each participant. The resulting β values were converted to t values (75), resulting in a whole-brain map of t values for each trial. The duration of the

epoch was 3 s for the decision phase and 2 s for the level of difficulty phase. As for the previous analyses, average background, WM and CSF signal, framewise displacement, six head motion regressors, and six aCompCor regressors, all obtained from fMRIprep, were entered as regressors of no interest. All regressors were convolved with the canonical hemodynamic response function. Multilevel modeling was conducted with custom R scripts in combination with the lme4 package for linear mixed-effects models (76) and the glmmlasso package for variable selection for generalized linear mixed models by L1-penalized estimation (37). fMRI analyses were conducted using custom Python scripts, which will be made publicly available.

Beta-series correlations. In order to further explore how the different areas resulting from the different contrasts described above interact with each other during decisions to cheat, we investigated the functional connectivity between these areas during the decision phase of the spot-the-difference task. To avoid the problem of activation-induced correlations we implemented beta-series correlations (39). We used the single-trial activations obtained as explained above by fitting a model that includes a separate regressor for each trial. We then correlated the parameter estimates from these regressors (the beta series) for honest decisions and cheated decisions separately between all of the regions found to be significantly related to our contrast of interest, in order to examine the degree to which they show similar trial-by-trial activations, as is expected when these regions were functionally connected. The beta-series model is particularly useful in event-related fMRI studies where the spacing between trials is relatively long (more than 8 to 10 s), which is the case in our paradigm (77). After obtaining the correlation matrix for each of the participants for honest and cheated decisions, we then also correlated the functional connectivity between each of the regions with the cheat count (individual differences in total cheating) in order to examine how functional connectivity differed for cheaters and more honest participants. To compare functional connectivity between honest and cheated decisions, correlations were transformed to z values using the Fisher r-to-z transformation. Significance was estimated by means of permutation testing where the cheat count was randomly shuffled at each iteration (n = 5,000). The resulting empirical P values were then corrected for multiple comparisons at FDR < 0.05.

Data Availability. Data, scripts, and images used in the task are available in Figshare (https://datarepository.eur.nl/articles/To_cheat_or_not_to_cheat_Cognitive_control_processes_override_our_moral_default/12287807).

ACKNOWLEDGMENTS. We gratefully acknowledge financial support from the Erasmus Research Institute of Management and the Dutch national e-infrastructure with support of the SURF Cooperative. In addition, we thank Dr. Hang-Yee Chan for advice and support throughout the project and Dr. Alex Genevsky for helpful comments on the final draft.

1. G. S. Becker, Crime and punishment: An economic approach. *J. Polit. Econ.* **76**, 169–217 (1968).
2. M. G. Allingham, A. Sandmo, Income tax evasion: A theoretical analysis. *J. Public Econ.* **1**, 323–338 (1972).
3. J. Gash, *The Great California Game: A Lovejoy Mystery*, (Penguin Books, 1992).
4. T. G. Seuntjens, M. Zeelenberg, N. van de Ven, S. M. Breugelmans, Greedy bastards: Testing the relationship between wanting more and unethical behavior. *Pers. Individ. Dif.* **138**, 147–156 (2019).
5. N. Abe, J. D. Greene, Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. *J. Neurosci.* **34**, 10564–10572 (2014).
6. E. Q. Campbell, The internalization of moral norms. *Sociometry* **27**, 391–412 (1964).
7. J. Henrich *et al*., In search of Homo economicus: Behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* **91**, 73–78 (2001).
8. E. Aronson, "A theory of cognitive dissonance: A current perspective" in *Advances in Experimental Social Psychology*, L. Berkowitz, Ed. (Academic Press, New York, 1969), Vol. 4, pp. 1–34.
9. R. F. Baumeister, "The self" in *Handbook of Social Psychology*, D. T. Gilbert, S. T. Fiske, G. Lindzey, Eds. (McGraw-Hill, New York, 1998), pp. 680–740.
10. D. J. Bem, "Self-perception theory" in *Advances in Experimental Social Psychology*, L. Berkowitz, Ed. (Academic Press, New York, 1972), Vol. 6, pp. 1–62.
11. R. Dhar, K. Wertenbroch, Self-signaling and the costs and benefits of temptation in consumer choice. *J. Mark. Res.* **49**, 15–25 (2012).
12. S. Berthoz, J. Grèzes, J. L. Armony, R. E. Passingham, R. J. Dolan, Affective response to one's own moral violations. *Neuroimage* **31**, 945–950 (2006).
13. N. Mazar, O. Amir, D. Ariely, The dishonesty of honest people: A theory of self-concept maintenance. *J. Mark. Res.* **45**, 633–644 (2008).
14. D. A. Gusnard, E. Akbudak, G. L. Shulman, M. E. Raichle, Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4259–4264 (2001).
15. H. Meffert, L. Blanken, K. S. Blair, S. F. White, J. R. Blair, The influence of valence and decision difficulty on self-referential processing. *Front. Hum. Neurosci.* **7**, 46 (2013).
16. M. van Buuren, T. E. Gladwin, B. B. Zandbelt, R. S. Kahn, M. Vink, Reduced functional coupling in the default-mode network during self-referential processing. *Hum. Brain Mapp.* **31**, 1117–1127 (2010).
17. F. Gino, M. E. Schweitzer, N. L. Mead, D. Ariely, Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organ. Behav. Hum. Decis. Process.* **115**, 191–203 (2011).
18. J. D. Greene, J. M. Paxton, Patterns of neural activity associated with honest and dishonest moral decisions. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12506–12511 (2009).
19. M. A. Maréchal, A. Cohn, G. Ugazio, C. C. Ruff, Increasing honesty in humans with noninvasive brain stimulation. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4360–4364 (2017).
20. N. L. Mead, R. F. Baumeister, F. Gino, M. E. Schweitzer, D. Ariely, Too tired to tell the truth: Self-control resource depletion and dishonesty. *J. Exp. Soc. Psychol.* **45**, 594–597 (2009).
21. N. C. Köbis, B. Verschuere, Y. Bereby-Meyer, D. Rand, S. Shalvi, Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspect. Psychol. Sci.* **14**, 778–796 (2019).
22. S. M. McClure, D. I. Laibson, G. Loewenstein, J. D. Cohen, Separate neural systems value immediate and delayed monetary rewards. *Science* **306**, 503–507 (2004).
23. C. M. Barnes, J. Schaubroeck, M. Huth, S. Ghummam, Lack of sleep and unethical conduct. *Organ. Behav. Hum. Decis. Process.* **115**, 169–180 (2011).
24. S. Shalvi, O. Eldar, Y. Bereby-Meyer, Honesty requires time (and lack of justifications). *Psychol. Sci.* **23**, 1264–1270 (2012).
25. K. Suchotzki, B. Verschuere, B. Van Bockstaele, G. Ben-Shakhar, G. Crombez, Lying takes time: A meta-analysis on reaction time measures of deception. *Psychol. Bull.* **143**, 428–453 (2017).
26. B. Verschuere, N. C. Köbis, Y. Bereby-Meyer, D. Rand, S. Shalvi, Taxing the brain to uncover lying? Meta-analyzing the effect of imposing cognitive load on the reaction time costs of lying. *J. Appl. Mem. Cognit.* **7**, 462–469 (2018).
27. V. Capraro, Does the truth come naturally? Time pressure increases honesty in one-shot deception games. *Econ. Lett.* **158**, 54–57 (2017).
28. V. Capraro, J. Schulz, D. G. Rand, Time pressure and honesty in a deception game. *J. Behav. Exp. Econ.* **79**, 93–99 (2019).
29. A. E. Van't Veer, M. Stel, I. Van Beest, Limited capacity to lie: Cognitive load interferes with being dishonest. *Judgm. Decis. Mak.* **9**, 199–206 (2014).
30. P. J. Gai, "Lies of bilingual consumers" in *Contextualized Consumers: Theories and Evidence on Consumer Ethics, Product Recommendations, and Self-Control*, (Erasmus University Rotterdam, Rotterdam, Netherlands, 2020), pp. 15–46.
31. U. Gneezy, A. Kajackaite, J. Sobel, Lying aversion and the size of the lie. *Am. Econ. Rev.* **108**, 419–453 (2018).
32. R. A. Poldrack, Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* **10**, 59–63 (2006).
33. T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, T. D. Wager, Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
34. K. Ballard, B. Knutson, Dissociable neural representations of future reward magnitude and delay during temporal discounting. *Neuroimage* **45**, 143–150 (2009).
35. B. Knutson, C. M. Adams, G. W. Fong, D. Hommer, Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci.* **21**, RC159 (2001).
36. S. Oldham *et al*., The anticipation and outcome phases of reward and loss processing: A neuroimaging meta-analysis of the monetary incentive delay task. *Hum. Brain Mapp.* **39**, 3398–3418 (2018).
37. A. Groll, G. Tutz, Variable selection for generalized linear mixed models by L 1-penalized estimation. *Stat. Comput.* **24**, 137–154 (2014).
38. R. Tibshirani, Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
39. J. Rissman, A. Gazzaley, M. D'Esposito, Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage* **23**, 752–763 (2004).
40. D. D. Cox, R. L. Savoy, Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* **19**, 261–270 (2003).
41. T. M. Mitchell *et al*., Learning to decode cognitive states from brain images. *Mach. Learn.* **57**, 145–175 (2004).
42. E. Vul, C. Harris, P. Winkielman, H. Pashler, Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* **4**, 274–290 (2009).
43. D. T. Welsh, L. D. Ordonez, The dark side of consecutive high performance goals: Linking goal setting, depletion, and unethical behavior. *Organ. Behav. Hum. Decis. Process.* **123**, 79–89 (2014).
44. C. M. MacLeod, Half a century of research on the Stroop effect: An integrative review. *Psychol. Bull.* **109**, 163–203 (1991).
45. B. A. Eriksen, C. W. Eriksen, Effects of noise letters upon the identification of a target letter in a nonsearch task. *Percept. Psychophys.* **16**, 143–149 (1974).
46. J. R. Simon, J. D. Wolf, Choice reaction time as a function of angular stimulus-response correspondence and age. *Ergonomics* **6**, 99–105 (1963).
47. P. Qin, G. Northoff, How is our self related to midline regions and the default-mode network? *Neuroimage* **57**, 1221–1233 (2011).
48. A. Vanhaudenhuyse *et al*., Two distinct neuronal networks mediate the awareness of environment and of self. *J. Cogn. Neurosci.* **23**, 570–578 (2011).
49. R. L. Carhart-Harris *et al*., The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Front. Hum. Neurosci.* **8**, 20 (2014).
50. H. L. Gallagher, C. D. Frith, Functional imaging of "theory of mind". *Trends Cogn. Sci.* **7**, 77–83 (2003).

51. A. Ciaramidaro *et al*., The intentional network: How the brain reads varieties of intentions. *Neuropsychologia* **45**, 3105–3113 (2007).
52. M. Schurz, J. Radua, M. Aichhorn, F. Richlan, J. Perner, Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neurosci. Biobehav. Rev.* **42**, 9–34 (2014).
53. R. L. Buckner, J. R. Andrews-Hanna, D. L. Schacter, The brain's default network: Anatomy, function, and relevance to disease. *Ann. N Y Acad. Sci.* **1124**, 1–38 (2008).
54. M. E. Raichle, A. Z. Snyder, A default mode of brain function: A brief history of an evolving idea. *Neuroimage* **37**, 1083–1090, discussion 1097–1099 (2007).
55. R. N. Spreng, R. A. Mar, A. S. Kim, The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *J. Cogn. Neurosci.* **21**, 489–510 (2009).
56. S. P. Speer, A. Smidts, M. A. Boksem, Individual differences in (dis)honesty are represented in the brain's functional connectivity: Robust out-of-sample prediction of cheating behavior. bioRxiv:10.1101/2020.05.12.091116 (14 May 2020).
57. J. D. Greene, L. E. Nystrom, A. D. Engell, J. M. Darley, J. D. Cohen, The neural bases of cognitive conflict and control in moral judgment. *Neuron* **44**, 389–400 (2004).
58. S. A. Spence *et al*., Behavioural and functional anatomical correlates of deception in humans. *Neuroreport* **12**, 2849–2853 (2001).
59. D. D. Langleben *et al*., Brain activity during simulated deception: An event-related functional magnetic resonance study. *Neuroimage* **15**, 727–732 (2002).
60. P. Gerlach, K. Teodorescu, R. Hertwig, The truth about lies: A meta-analysis on dishonest behavior. *Psychol. Bull.* **145**, 1–44 (2019).
61. B. E. Hilbig, I. Thielmann, Does everyone have a price? On the role of payoff magnitude for ethical decision making. *Cognition* **163**, 15–25 (2017).
62. J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, J. D. Cohen, An fMRI investigation of emotional engagement in moral judgment. *Science* **293**, 2105–2108 (2001).
63. N. Lisofsky, P. Kazzer, H. R. Heekeren, K. Prehn, Investigating socio-cognitive processes in deception: A quantitative meta-analysis of neuroimaging studies. *Neuropsychologia* **61**, 113–122 (2014).
64. M. Bahnemann, I. Dziobek, K. Prehn, I. Wolf, H. R. Heekeren, Sociotopy in the temporoparietal cortex: Common versus distinct processes. *Soc. Cogn. Affect. Neurosci.* **5**, 48–58 (2010).
65. K. Gorgolewski *et al*., Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.*, 10.3389/fninf.2011.00013 (2011).
66. O. Esteban *et al*., fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
67. O. Esteban *et al*., Surface-driven registration method for the structure-informed segmentation of diffusion MR images. *Neuroimage* **139**, 450–461 (2016).
68. C. Caballero-Gaudes, R. C. Reynolds, Methods for cleaning the BOLD fMRI signal. *Neuroimage* **154**, 128–149 (2017).
69. N. J. Tustison *et al*., N4ITK: Improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).
70. S. M. Smith, Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**, 143–155 (2002).
71. R. W. Cox, J. S. Hyde, Software tools for analysis and visualization of fMRI data. *NMR Biomed.* **10**, 171–178 (1997).
72. J. M. Treiber *et al*., Characterization and correction of geometric distortions in 814 diffusion weighted images. *PLoS One* **11**, e0152472 (2016).
73. X. Y. Liu, J. Wu, Z. H. Zhou, Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. B Cybern.* **39**, 539–550 (2009).
74. J. A. Mumford, B. O. Turner, F. G. Ashby, R. A. Poldrack, Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage* **59**, 2636–2643 (2012).
75. M. Misaki, Y. Kim, P. A. Bandettini, N. Kriegeskorte, Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage* **53**, 103–118 (2010).
76. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. arXiv:1406.5823 (23 June 2014).
77. R. A. Poldrack, J. A. Mumford, T. E. Nichols, *Handbook of Functional MRI Data Analysis*, (Cambridge University Press, 2011).