Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Medial prefrontal cortex encodes and regulates trait honesty

Ce Mo

moce3@mail.sysu.edu.cn

Sun Yat-Sen University https://orcid.org/0000-0001-6666-9282

Yuan Gao

Lei Mo

Edmund Derrington

Claude Bernard University Lyon 1

Jean-Claude Dreher

Article

Keywords: honesty, motivational conflict, mPFC, tDCS, fMRI, psychometric model

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11213380/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Medial prefrontal cortex encodes and regulates trait honesty

Yuan Gao[1,#], Ce Mo[2,#,*], Lei Mo[1,*], Edmund Derrington[3], Jean-Claude Dreher[3]

[1]Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, 510630, P. R. China

[2]Department of Psychology, Sun Yat-Sen University, Guangzhou, 510006, P. R. China

[3]Neuroeconomics, Reward and Decision Making Group, Center for Cognitive Neurosciences, CNRS, University of Lyon 1, Lyon, 69675, France.

[#] Equal contribution

[*] Correspondence: moce3@mail.sysu.edu.cn, molei@m.scnu.edu.cn

Data availability statement

The data and code supporting the findings of this study are available at https://osf.io/r7dky/overview?view_only=6b3ea7fa2a594ba18f3172ed455fc142

Funding statement

This work was supported by grants from the MOE Major Special Project of Philosophical and Social Sciences Research (2025JZDZ024), the MOE Project of Key Research Institute of Humanities and Social Sciences in Universities (22JJD190007), and the Major Planning Project of Philosophy and Social Science of Guangdong Province (GD23ZD17) to L.M., Social Science Foundation of Guangdong Province, China (GD24YXL03) and Basic and Applied Research Foundation of Guangzhou, China (SL2023A04J00644) to C.M., IDEXLYON from Université de Lyon (project INDEPTH) within the Programme Investissements d'Avenir (ANR-16-IDEX-0005) and of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National

Research Agency. This work was also supported by grants from the Agence Nationale pour la Recherche to J-C D. (ANR-24-CE37-4261-01).

Acknowledgements

We thank all participants for taking part in the study.

Author contributions

Y.G., C.M., L.M. and J-C. D. designed the experiment. Y.G. collected the data. Y.G. and C.M. conducted data analyses. Y.G. wrote the manuscript. Y.G., C.M., L.M., E.D. and J-C. D. reviewed and edited the manuscript.

Competing interests statement

The authors declare no competing interests.

Ethical approval

This study was approved by the Ethics Committee of South China Normal University (Approval No. SCNU-PSY-2022-117).

Abstract

Honesty varies substantially across individuals, reflecting a trade-off between potential material gains and the moral costs of dishonest behavior. However, how such trait-level variability is represented in the brain and causally shapes dishonest behavior remain unclear. Here, we combined psychometric modeling, functional MRI and transcranial direct current stimulation (tDCS) to investigate the neural representation of trait honesty and its causal link to (dis)honest behavior. We quantified trait honesty as an individual-specific “honesty threshold”, the balance point between potential dishonest gains and moral costs of cheating. Higher honesty threshold predicted lower cheating rates and reduced sensitivity to increasing potential dishonest gains. Using fMRI, we show that individual differences in honesty threshold are encoded in the medial prefrontal cortex (mPFC), while motivational conflict during dishonest decision-making is tracked by the superior parietal cortex (SPC). Effective connectivity analyses reveal that mPFC modulates SPC activity, linking trait representation to conflict processing. Critically, increasing mPFC excitability via tDCS causally elevates honesty thresholds and reduces the rate at which cheating frequency

increased with potential dishonest gains. Together, these findings identify the mPFC as a neural substrate of trait-like honesty and establish a circuit mechanism through which trait representations regulate dishonest behavior under incentive conflict.

Keywords: honesty, motivational conflict, mPFC, tDCS, fMRI, psychometric model

Introduction

Honesty is a fundamental component of human social behavior and is widely valued across societies [1, 2]. However, individuals vary substantially in their propensity to behave honestly, particularly in situations where self-interest is at stake [3]. In such contexts, decisions reflect a trade-off between the desire for material gain and the aversion to the moral cost of dishonest behavior [4-9]. Converging evidence from neuroimaging and brain stimulation studies implicates a distributed neural network including the medial and dorsolateral prefrontal cortex, the anterior cingulate cortex, and the temporoparietal junction (TPJ) in (dis)honest decision-making [9-15]. Functional connectivity between these brain regions further predicts individual differences in dishonest behavior [16-18]. Beyond these neuroimaging findings, non-invasive brain stimulation and lesion studies revealed the causal influence of the prefrontal regions, most notably dlPFC, on (dis)honest behavior [7, 19, 20]. Despite these advances, it remains unclear how stable individual differences in honesty are represented in the brain and whether such representations exert a causal influence on behavior.

A central challenge in addressing this question is the lack of a quantitative, incentive-sensitive measure of trait honesty. Previous studies have often relied on binary behavioral classifications or self-report measures[18,43], which are limited in their ability to capture continuous individual differences and are susceptible to strategic behavior and social desirability biases[44-46]. Moreover, because the propensity to engage in (dis)honest behavior depends on the magnitude of potential gains[6,22], estimates derived from a single incentive context may confound stable trait differences with context-dependent strategies. For example, ‘disguised cheaters’ may be mischaracterized as relatively honest if they cheat strategically in high-reward trials but restrain themselves in low-reward trials. Thus, an unbiased characterization of trait honesty requires a framework that systematically samples across incentive

magnitudes and that formalizes individual differences in honesty as a continuous parameter that generalizes across contexts.

Here, we address this challenge by adopting a psychometric modeling approach to quantify trait honesty as an individual-specific "honesty threshold"—the level of potential dishonest gain at which an individual is equally likely to cheat or remain honest. This formulation captures a stable balance between material incentives and adherence to honesty norms, providing a continuous measure that can be mapped onto both behavior and neural activity. We combined this approach in a multimodal causal framework with functional magnetic resonance imaging (fMRI) to identify brain regions encoding individual differences in honesty threshold, and with non-invasive brain stimulation to test whether such representations causally influence (dis)honest behavior.

We designed a new version of a prominent spontaneous cheating task (i.e. the coin-flipping task[11]) in which the participant's true guess and self-report were inconspicuously recorded and compared in each trial, thus allowing for an unequivocal identification of dishonest behaviors at the trial level. Similar to the classical version of this paradigm, each trial was associated with a stake. Participants won the stake when their self-report matched the coin-flip outcome, or lost the same amount of money when they reported a failed prediction. To manipulate the incentive magnitude, we systematically varied the trial-wise stakes over a wide range of monetary amounts. Using a psychometric approach[23, 24], we parameterized each participant's trait honesty as 'honesty threshold', defined as the critical stake amount marking the balance point between dishonest gain and adherence to honesty norms. We show that honesty threshold predicts both overall cheating frequency and sensitivity to increasing potential dishonest gains. At the neural level, individual differences in honesty threshold are encoded in the medial prefrontal cortex (mPFC), a key

region involved in self-referential processing and value integration[31-36], whereas motivational conflict during dishonest decision-making is tracked by the superior parietal cortex. Effective connectivity analyses further reveal that mPFC modulates parietal conflict signals, linking trait-level representation to decision processes. Critically, increasing mPFC excitability via transcranial direct current stimulation causally elevates honesty thresholds and reduces cheating behavior[19]. Together, these findings highlight the central role of mPFC in representing a trait-like honesty parameter and regulating (dis)honest behavior, and establish a circuit mechanism through which it regulates dishonest behavior.

Results

fMRI experiment

To identify (dis)honest behavior associated with different incentive magnitudes, a new version of the coin-flipping task was employed with a meticulously designed cover story (see **Supplementary information** for the complete version of participant instruction). Before the fMRI scanning, participants were told that the study investigated the influence of privacy and mental efforts on intuitive hunches, and that they would be assigned to either a 'supervised group' or an 'unsupervised group'. They were always 'assigned' to the 'unsupervised group' on site by a scripted lottery program. On each trial, participants indicated their guess of the outcome of a random coin-flip with a corresponding key press, and then recalled their guess after the coin-flip (**Figure 1a**). To reduce the participants' suspicion, they were also told that the guesses of the 'supervised group' participants were recorded and visible, while their own key presses were completely private and served only to match the response procedure between the two groups. Furthermore, the participants were informed that a stake was associated with each trial to motivate mental effort. Based on whether their recalled guess matched the coin-flip outcome, they either obtained or lost the corresponding amount of money. This scheme was designed to incentivize

cheating in participants, as in Greene and Paxton [11]. Critically, unlike the original version of the task, participants unknowingly provided a verifiable record of their initial guesses on every trial. Moreover, the coin-flip outcomes matched participants' initial guesses on some of the trials (i.e., Match trials) but differed from them on the other trials (i.e., Mismatch trials). Hence, stake amount in the Mismatch trials indexed the magnitude of potential dishonest gain. Cheating was identified on the Mismatch trials when a participant's recalled guess matched the displayed outcome but differed from the initial guess.

Behavioral and neuroimaging data were collected simultaneously inside the fMRI scanner. After scanning, all participants underwent a *post-hoc* debriefing session in which they were interviewed individually to assess the effectiveness of the cover story and their confidence in privacy during the experiment. Five participants expressed doubts about the study's objective, saying that they suspected it "might have something to do with honesty/integrity, tendency of lying or desire for money in private settings". One participant requested termination of the scanning. Data from these participants were excluded from further analysis. All remaining participants were confident in the privacy of their decisions (**Figure 1b** left).

Honesty threshold provides a valid measure of trait honesty

For each participant, we computed the percentage of mismatch trials on which they cheated and the mean decision time at each stake amount. We found that cheating frequency increased gradually with stake amount ($F_{(5,115)} = 70.63$, $p < 0.001$) in a manner that was highly reminiscent of typical psychometric functions (**Figure 2a**). We therefore characterized the relationship between cheating frequency and stake amount using a well-established computational modeling approach in psychophysics. The upper bound of the model function represented the degree to which the participants preserved their honesty despite the increasingly tempting potential dishonest gains. The lower bound

represented the participants' default inclination to cheat (even in absence of potential dishonest gains). The shape of the model function was governed by two parameters, which jointly determined the pattern of how cheating frequency increased with incentive magnitude (i.e. stake amount). To evaluate the model's goodness of fit, we calculated the model deviance for each participant following the standard psychometric modeling procedure. Model deviance takes values from 0 (no residual error between model prediction and the observed data) to infinity (the observed data are impossible given the model predictions). A Monte-Carlo simulation was used to obtain the distribution of deviance that one would expect from a participant whose behavioral responses conformed to the model prediction [23, 24]. We found that the empirical deviance was located well within the two-sided 95% confidence interval of the simulated deviance distribution (**Figure 2b**), which was consistently observed across individual participants (**Figure S1**). Further, we compared the goodness of fit of the psychometric model against a simpler alternative (i.e. a linear model) by computing both the Akaike Information Criterion (AIC) corresponding to the two models. We found that our model provided a better fit to the data in comparison to a simpler linear model (median ΔAIC = -150.19, bootstrapped 95% C.I. = [-168.57, -129.74], Wilcoxon's signed rank test, $p < 0.001$; **Figure 2c**). These findings indicated that our psychometric model accurately captured the participants' stake-dependent cheating behavior, even at the individual level.

Having demonstrated the psychometric model's fitness, we defined honesty threshold as the stake amount corresponding to the midpoint of the fitted psychometric model function (i.e., halfway between the lower and upper bounds). Conceptually, honesty threshold signified the point of indifference between potential dishonest gains and the moral cost of cheating, thereby serving as a trait-like honesty parameter. This conceptualization yields two key validity checks. First, individuals with a higher honesty threshold should exhibit a shallow rise in cheating frequency with increasing potential dishonest gains.

Second, consistent with the self-concept maintenance theory [8], the level of motivational conflict should depend jointly on the amount of potential dishonest gain and one's honesty-related self-concept constraints. Conflict should be low at very small or very large stake amounts, but peak when the stake amount matches the honesty threshold. The heightened conflict should in turn prolong decision time.

For the first check, we quantified the rate at which cheating frequency increased with potential dishonest gains using a multi-level regression model for repeated measures [25, 26]. In this model, trial-wise stake amount was entered as a level-1 predictor, and each participant's honesty threshold (a subject-level parameter) was entered as a level-2 predictor. Cheating frequency across stake amounts served as the dependent variable. The level-1 coefficient captured the rate at which cheating frequency increased with stake amount (i.e. the stake-cheating slope). As a preliminary examination, we first verified that individuals with higher honesty threshold made fewer dishonest decisions (between-subject effect for honesty threshold: b = -0.35, S.E. = 0.08, t_{24} = -4.44, p < 0.001). More importantly, consistent with our expectation, the stake-cheating slope was dependent on honesty threshold, as indicated by a significant cross-level interaction effect between stake amount and honesty threshold (b = -0.10, S.E. = 0.05, t_{120} = -2.12, p = 0.036). Specifically, while there was a sharp rise in cheating frequency with increasing stake amount for individuals with low honesty thresholds (simple slope coefficient: 0.76, S.E. = 0.07, p < 0.001), individuals with high honesty thresholds exhibited a reduced rate of increase (simple slope coefficient: 0.56, S.E. = 0.07, p < 0.001) (**Figure 2d**). For the second check, we analyzed the relationship between the level of motivational conflict and the decision time using the same regression model approach. For each participant, we operationalized motivational conflict as the absolute distance between stake and the honesty threshold, with smaller distances indicating higher conflict. The operationalized motivational conflict index served

as a fixed-effect predictor of the regression model, and the mean decision time for each non-zero stake amount served as the dependent variable. As expected, we found that decision time decreased with the increasing absolute stake-threshold distance (b = -0.23, S.E. = 0.04, t_{97} = -5.93, p < 0.001, **Figure 2e**). Together, the results of the two validity checks provided converging support for honesty threshold as a valid measurement of trait honesty.

Stake amount is encoded in reward-related regions

As a preliminary step of the fMRI data analysis, we first identified the brain regions involved in encoding stake amount. We reasoned that the neural activation of these regions should track the trial-wise stake amount, even in Match trials where no trade-off between honesty and dishonest gains was required. We further predicted that these brain regions should be located within the canonical 'reward circuit', given its well-documented involvement in processing reward information [27-29]. To locate these brain regions, a general linear model (GLM 1) was built for each individual subject. Specifically, Mismatch trials and Match trials were modeled together with a single regressor, and trial-wise stake amount was entered as a parametric modulator. Moreover, the ineffective trials (i.e. trials with no or implausible responses) were modeled with an additional regressor (see **Methods** for details). The subject-level contrast maps corresponding to the parametric modulator were submitted to a one-sample *t*-test at the group level to identify the brain regions whose neural activation was parametrically modulated by stake amount. We found that the blood oxygen level-dependent (BOLD) signal intensity in multiple brain regions of the reward-related network increased with the trial-wise stake amount (**Figure 3a,** see **Table 1** for brain region details). In particular, this effect was strongest in two central nodes of the canonical reward circuit, namely nucleus accumbens (NAcc) and dorsal anterior cingulate cortex (dACC). This pattern is consistent with previous works on reward processing [12, 28, 30], which serves as a sanity check for the imaging data.

mPFC encodes individual differences in honesty threshold

Having parameterized trait honesty as the honesty threshold, we asked whether and where honesty threshold is represented in the brain. We reasoned that the neural representation of honesty threshold should be engaged when participants needed to evaluate the trade-off between dishonest gains and moral costs, which corresponded to the Mismatch trials with non-zero stakes in our design. Most importantly, the brain regions containing this representation should encode the individual differences in honesty threshold across participants. Because trait honesty is closely tied to a person's self-concept, we further predicted that honesty threshold should be represented in the key regions within the self-referential processing network. To test this, we built a subject-level general linear model (GLM 2) that included separate regressors for Match trials, Mismatch trials with zero stake, Mismatch trials with non-zero stakes, and ineffective trials (see **Methods** for details). Because honesty threshold is conceptualized as a stable individual-level trait-like parameter across incentive magnitudes, we computed the activation contrast between the Mismatch trials with non-zero stakes and the Mismatch trials with zero stake for each participant. At the group level, the brain regions representing honesty threshold were located by correlating the subject-level contrast estimates in each voxel with the individual honesty threshold across all participants. Although the subject-level contrast might in principle include stake-related responses, we controlled for this potential confound by applying an exclusive mask to the group-level results. This mask was derived from the group-level map corresponding to the brain regions that were parametrically modulated by stake amount (obtained from GLM 1), which was liberally thresholded ($p < 0.005$, uncorrected) to ensure better coverage of stake-related activations. As a result, we found that the BOLD signal intensity of right medial prefrontal cortex (mPFC) positively correlated with individual honesty thresholds (**Figure 3b,** see **Table 2** for brain region details), suggesting that mPFC represents a trait-like

honesty parameter.

If honesty threshold is indeed represented in mPFC, then mPFC activation should relate to dishonest behavior in a manner that parallels the behavioral signature of honesty threshold. To test this, we defined the supra-threshold cluster in mPFC as a region of interest (ROI), and extracted each subject's mean contrast estimate within the ROI as a measure of mPFC activation. We then conducted the same multi-level regression analysis, with the original honesty threshold predictor replaced by the mPFC activation measure. This analysis reproduced the same pattern observed behaviorally (**Figure 3d**). Specifically, lower cheating frequency was associated with stronger mPFC activations across participants (between-subject effect: b = -0.23, S.E. = 0.09, t_{24} = -2.41, p = 0.024). Moreover, the stake-cheating relationship was constrained by mPFC activation (mPFC activation × stake amount interaction: b = -0.10, S.E. = 0.05, t_{120} = -1.95, p = 0.044), where cheating frequency rose with stake amount at a slower rate in participants with higher mPFC activation (simple slope coefficient: 0.56, S.E. = 0.07, p < 0.001) relative to those with lower mPFC activation (simple slope coefficient: 0.76, S.E. = 0.07, p < 0.001). Together, these findings demonstrated a close correspondence between honesty threshold and mPFC activation, further supporting that a trait-like honesty parameter is represented in mPFC, a canonical hub of the self-referential network [31-36]

SPC signals motivational conflict

Next, we investigated the brain region(s) signaling the motivational conflict during (dis)honest decision making in incentive contexts. We reasoned that these regions should be engaged in Mismatch trials with non-zero stakes where participants had to weigh dishonest gains against the moral cost of cheating. Moreover, neural activation of these regions should track the operationalized motivational conflict index (i.e. the absolute stake-threshold distance) on a trial-

by-trial basis. To locate these regions, a parametric modulator corresponding to the absolute stake-threshold distance was included in GLM 2 (see **Methods** for details). Individual contrast maps corresponding to the parametric modulator were entered to a group-level one-sample *t*-test to localize the brain regions whose neural activation scaled with the absolute stake-threshold distance. To control for potential stake-related activation, we employed the same exclusive masking approach to the group results as described earlier. Significant effect of the parametric modulator was found in bilateral superior parietal cortex (SPC), suggesting that neural activity in SPC tracks motivational conflict across incentive magnitudes (**Figure 3c,** see **Table 3** for brain region details).

If SPC indeed signals motivational conflict, then stronger SPC activation should be associated with longer decision time. To test this, we estimated the SPC activation corresponding to each incentive magnitude for each participant using a separate subject-level GLM (GLM 3). In GLM 3, Mismatch trials with different stake amounts were modeled with separate regressors (see **Methods** for details). We then defined the supra-threshold clusters in bilateral SPC as an ROI, and computed the ROI's mean beta estimate corresponding to each non-zero stake amount. Similar to the behavioral data analysis, a linear mixed model was used to investigate the relationship between the ROI's mean beta estimate and the decision time. We found that the ROI's mean beta estimates were positively related to decision time (b = 0.29, S.E. = 0.09, t_{119} = 3.06, p = 0.003) (**Figure 3e**). This finding further supported that the motivational conflict associated with making (dis)honest decisions is signaled by SPC.

mPFC modulates SPC via direct effective connectivity

The functional localization results raise the question of whether and how mPFC and SPC interact to support (dis)honest decision-making in incentive contexts. A dynamic causal model (DCM) analysis was employed to investigate the functional interplay between mPFC and SPC. Because this decision-making

process starts with encoding of monetary information, stake amount was entered as the extrinsic driving input to dACC, the functional cluster with the strongest stake-tracking effect in GLM 1. In a similar vein, left SPC was chosen as the representative node corresponding to motivational conflict. Because the neural signals of honesty threshold are expected to be engaged when participants weigh potential dishonest gains against moral costs, the dACC → mPFC connection should be modulated by the presence of potential dishonest gains (i.e. Mismatch trials with non-zero stakes). Moreover, because motivational conflict depends jointly on both honesty threshold and stake amount, we considered the cases in which SPC received inputs from both mPFC and dACC, thereby allowing conflict-related activity to be influenced by both sources.

Based on the rationale above, we consider four possibilities concerning the functional interplay between mPFC and SPC. First, mPFC directly influences SPC, in which case the amount of potential dishonest gains modulates the mPFC → SPC connectivity (direct pathway). Moreover, the connectivity between mPFC and dACC could be either uni-directional (model 1) or bi-directional (model 2). Second, mPFC exerts its influence on SPC with dACC as its proxy (indirect pathway), in which case the amount of potential dishonest gains modulated both the mPFC → dACC connectivity and the dACC → SPC connectivity (model 3). Third, mPFC influences SPC via both direct and indirect pathways (model 4). Fourth, there is no modulatory effect on the connectivity between mPFC and SPC (models 5 & 6). These candidate models (see **Figure 4a** for a complete illustration) were fit to each participant's data. Using a hierarchical Bayesian approach, we identified the winning model by comparing the exceedance probability of the candidate models. The exceedance probability corresponds to the posterior probability that a given model is more likely than any other candidate model.

We found that our data were best explained by the model featuring a direct mPFC→SPC connectivity and bi-directional coupling between mPFC and dACC (model 2, **Figure 4b&c**). Specifically, the presence of potential dishonest gains significantly increased the dACC → mPFC connectivity (one-sample *t* test: $t_{23} = 3.195$, $p = 0.004$). Moreover, the effective connectivity from mPFC to SPC was significantly increased for the largest stake amount in comparison to that for the smallest stake amount (paired-sample *t* test: $t_{23} = 2.801$, $p = 0.010$), consistent with a direct influence of mPFC on SPC. Together, these observations unveiled a neural circuit underlying (dis)honest decision-making in incentive contexts, and suggested that neural signals of motivational conflict are directly modulated by the neural representation of the trait-like honesty parameter.

tDCS experiment

While the preceding results demonstrated that honesty threshold is represented by mPFC, a critical step beyond the fMRI findings is to establish a direct causal link between the representation of this trait-like honesty parameter and (dis)honest behavior. In the second experiment, we investigated the causal influence of mPFC activity on (dis)honest behavior in another large, independent sample of participants using transcranial direct current stimulation (tDCS). tDCS is a non-invasive method for localized modulation of neural activity in healthy human brain by applying weak electric currents to the scalp. Anodal stimulation exogenously enhances neural excitability of the target region, whereas sham stimulation produces no effective changes in neural excitability. Since tDCS has long-lasting after-effects [37-40], a within-subject design in tDCS experiments requires multiple sessions with a sufficiently long time interval between consecutive sessions, such that the after-effects of the preceding session could subside without interfering with those of the current session. However, because the contextualization procedure involved briefing the participants and above all, the presence of a 'supervised group participant',

it was extremely difficult, if not impossible, to ensure that the contextualization remained equally effective in different sessions. We therefore chose a between-subject design over a within-subject design.

Participants were randomly assigned to two independent stimulation groups (*N* = 50 in each group). In a single session, we administered offline anodal or sham tDCS at an intensity of 2mA for 20 minutes via a 4×1 ring electrode configuration, with F2 as the central node surrounded by four equidistant nodes (AFz, AF4, FCz, FC4) (**Figure 5a** top panel). Notably, cathodal tDCS was not performed because it has been reported to induce less stable cognitive and behavioral effects than anodal tDCS [19, 41]. As shown in the bottom panel of **Figure 5a**, the electric stimulation was delivered over the immediate vicinity of the right mPFC ROI identified from the fMRI findings. After stimulation, both groups underwent the same procedure and *post-hoc* debriefing as in the fMRI experiment. A small number of participants who reported suspicion about the study's objective, reported being unaware of the possibility of cheating for money, or requested termination were excluded from further analyses. No participants expressed concerns about their privacy during the experiment (**Figure 1b** middle & right). In addition, a safety questionnaire was administered to the participants in which they rated their subjective discomfort in various aspects (see **Methods** for details). None of the rating scores significantly differed between the two groups (independent two-sample *t*-test, all *p*s > 0.1). No side effects or discomfort were reported by any subject in either group.

Enhancement of mPFC excitability increases honesty threshold and reduces cheating

We first evaluated whether our model-based approach for estimating honesty threshold could be generalized to the larger tDCS sample. Similar to the fMRI experiment, we found that cheating frequency increased non-linearly with stake amount (**Figure 5b&c** upper panels) in both stimulation groups, which was

better captured by the psychometric model than by the linear model (anodal: median ΔAIC = -98.429, bootstrap 95%-C.I. = [-109.536, -73.123], sham: median ΔAIC = -134.444, bootstrap 95%-C.I. = [-149.550, -111.920], Wilcoxon's signed rank test: both *p*s **<** 0.001, **Figure 5d**). The agreement between the data and the psychometric model was corroborated by Monte Carlo simulations (**Figure 5b&c** bottom panels, see also **Figure S2&S3** for individual results). These findings demonstrated the robustness of our psychometric modeling approach, supporting the validity of the resulting honesty threshold estimates. We then asked whether enhancing mPFC excitability could causally modulate the representation of honesty threshold, in which case anodal stimulation should shift the honesty threshold estimates. To test this, we obtained the honesty threshold from the fitted psychometric model for each subject, and compared it between the two stimulation groups. As a result, honesty threshold was significantly higher in the anodal group (mean threshold ± S.E.M.: anodal: 7.09 ± 0.51, sham: 5.13 ± 0.55, two sample *t*-test: t_{98} = 2.621, p = 0.01). This finding indicated that modulation of mPFC activity shifted the honesty threshold estimates, which provided direct causal evidence for the critical role of mPFC in representing a trait-like honesty parameter.

Having validated the stimulation-induced change in honesty threshold, we next investigated stimulation-related changes in behavior and motivational conflict. To this end, we conducted the same multi-level regression analysis with stake amount as the level-1 predictor, and stimulation group (anodal vs. sham) as the level-2 predictor. We found that while both groups exhibited a significant rise in cheating frequency with increasing stake amount (both *p*s < 0.001), they differed significantly in the rate at which cheating frequency increased with potential dishonest gains (stimulation group × stake amount interaction: b = 0.14, S.E. = 0.04, $t_{499.10}$ = 3.61, p < 0.001, **Figure 5e**). In contrast to the sham group (simple slope coefficient: 0.45, S.E. = 0.03, p < 0.001), we observed in the anodal group a 'desensitized' pattern in which the rise in cheating frequency

493 with stake amount was less pronounced (simple slope coefficient: 0.30, S.E. =
494 0.02, $p < 0.001$). Specifically, anodal stimulation elicited an evident reduction in
495 cheating frequency across stake amounts ($F_{(5,490)} = 6.03$, $p = 0.002$), even for
496 large stakes. Moreover, the 'desensitization' to potential dishonest gains was
497 also echoed by a stimulation-induced change in the effect of motivational
498 conflict on decision time (**Figure 5f**). A larger absolute stake–threshold distance
499 (i.e., lower motivational conflict) was associated with shorter decision times in
500 the sham group (b = -0.06, S.E. = 0.02, $t_{201.29}$ = -2.88, p = 0.004), as we
501 observed for the participants in the fMRI experiment. In contrast, the stake–
502 threshold distance was not significantly related to decision time in the anodal
503 group (b = -0.03, S.E. = 0.03, $t_{200.80}$ = -1.35, p = 0.179), suggesting that
504 motivational conflict was less strongly reflected in decision times under anodal
505 stimulation. Together, these findings showed that anodal stimulation over the
506 mPFC region caused reduction of cheating behavior and attenuation of the
507 motivational conflict.
508
509 **Discussion**
510 In the present study, we combined a psychometric approach, functional MRI
511 and non-invasive brain stimulation in a multi-modal causal framework to
512 investigate the neural mechanism underlying (dis)honest decision-making. By
513 parameterizing trait honesty as an individual-specific "honesty threshold"
514 (amount of potential dishonest gain balanced against adherence to honesty
515 norms), we provide a psychometric quantitative framework that links behavior
516 across incentive contexts to a stable latent variable. We show that this trait-like
517 parameter is represented in the mPFC, that motivational conflict during
518 dishonest decision-making is signaled in the SPC, and that enhancing mPFC
519 excitability reduces dishonest behaviors and the effects of motivational conflict
520 in decision time. Furthermore, the two regions composed an interacting cortical
521 circuit in which SPC was modulated by mPFC via direct effective connectivity.
522 Together, these findings identify a neural substrate for trait honesty and

establish a circuit mechanism through which it regulates (dis)honest behavior.

Our psychometric modeling approach for estimating honesty threshold addresses a central limitation in prior work on honesty. Effective characterization of trait honesty across different individuals has been a core challenge in identifying the neural correlates of honesty. Earlier studies have often categorized people into qualitatively distinctive groups (e.g. honest people vs. cheaters) using spontaneous cheating tasks. In such tasks, dishonesty is inferred from overstated performance in unsupervised settings without access to veridical trial outcomes [10-12, 42]. Although such approaches can afford high sensitivity to experimental manipulations, they constrain the generalizability of the experimental findings to the real-world scenario in which honesty is not dichotomized. Moreover, they are also vulnerable to the concern of 'disguised cheaters' who cheat strategically for high rewards but restrained themselves when rewards are low. Another approach is to assess trait honesty via self-report measures [18, 43]. However, self-report measures are often biased by social desirability, impression management and limited introspective access to one's own moral motivations [44-46]. Consequently, self-report measures may fail to reliably capture the latent propensity that gives rise to dishonest behavior, particularly across varying incentive contexts. Inspired by psychophysics, we addressed this challenge by parameterizing trait honesty as honesty threshold (i.e. the amount of potential dishonest gain at which an individual is equally likely to cheat or to remain honest). Our framework thus provides a continuous quantitative measure of trait honesty that is less susceptible to these limitations.

Our results speak directly to ongoing debate regarding the role of cognitive control in honesty. A prominent account proposes that cognitive control is required to override selfish impulses and promote honest behavior[11]. However, recent evidence suggests a more nuanced picture: cognitive control may instead facilitate deviations from an individual's moral default, promoting honesty in habitual cheaters but enabling dishonesty in generally honest individuals[12]. Consistent with this framework, activity in cognitive control

regions such as the ACC and dlPFC has been shown to support both honest and dishonest behavior depending on baseline tendencies[12]. Our findings extend this perspective by identifying a trait-like parameter that formalizes this "moral default." Rather than classifying individuals categorically as honest or dishonest, the honesty threshold captures a continuous balance between material incentives and moral costs. Within this framework, cognitive control processes may act on top of this baseline parameter, facilitating shifts in behavior relative to an individual's position along the threshold continuum. This provides a quantitative bridge between prior accounts of moral default and trial-by-trial decision processes.

At the neural level, our findings provide novel and more comprehensive insights into the neural correlates of honesty. We showed that honesty threshold is represented in mPFC, a central hub of the cortical midline system for self-referential processing [47-49] and a key region closely related to altruism, cooperative behavior and self-image maintenance [32, 50-52]. Other studies have identified the dlPFC and ACC, as a important nodes to resolve motivational conflict in (dis)honest decision-making. For example, dlPFC lesions reduce the frequency of honest behaviors[7], and boosting the neural excitability in this region increases honest reporting in a die-roll paradigm [19]. Our findings hence suggest a functional dissociation between mPFC and regions associated with cognitive control. While the ACC and dlPFC may be responsible for resolving motivation conflict, the mPFC may integrate the information of dishonest gain with the social norms of honesty.

We also identify SPC as supporting a domain-general computation of motivational conflict during (dis)honest decision-making rather than encoding moral values *per se*. By operationalizing conflict as the distance between incentive magnitude and an individual's honesty threshold, we show that SPC activity tracks trial-by-trial fluctuations in this conflict signal, with maximal engagement when potential gains are closely balanced against moral costs. This pattern, together with the observed relationship between SPC activation

and decision time, suggests that SPC implements a value-comparison process that weighs potential gains against moral costs [53,54]. Importantly, effective connectivity analyses indicate that mPFC modulates SPC activity, linking trait-level representations of honesty to conflict-related processing. This circuit architecture provides a mechanistic account of how stable individual differences in honesty threshold influence moment-to-moment decisions.

Beyond correlational evidence, our tDCS results establish a causal role for mPFC in regulating dishonest behavior. Enhancing mPFC excitability increased honesty thresholds and reduced both overall cheating and sensitivity to incentive magnitude. These findings extend prior work demonstrating causal effects of prefrontal stimulation on honesty[19] by showing that modulation of mPFC specifically alters a trait-like parameter governing behavior. Moreover, the attenuation of conflict-related effects on decision time suggests that increasing mPFC excitability may shift the balance between competing motivations, thereby reducing the need for conflict resolution during decision-making.

Several limitations should be noted. First, our paradigm focuses on self-interested dishonesty in private settings and does not address forms of honesty involving social or public consequences, which may engage additional mechanisms[15,50]. Second, the psychometric framework assumes a monotonic relationship between incentive magnitude and cheating behavior. While behavioral patterns of most participants were in line with this monotonicity, it is important to recognize the possibility of individuals whose behaviors are not significantly influenced by stake amount (i.e., a flat profile of cheating frequency), which calls for further methodological refinements. In addition, the influence of audience also remains to be investigated [15, 50].

In summary, we provide a quantitative and mechanistic account of trait honesty by linking a psychometrically defined parameter to mPFC representation and causal influence of this region on honesty threshold and reduction of cheating behaviors. Our findings identify mPFC as a core locus of trait honesty, delineate

its interaction with parietal mechanisms of motivational conflict, and offer a general framework for a mechanistic understanding of the neural circuit underlying (dis)honest decision-making.

Methods

Participants

159 freshmen from the first author's affiliation were initially recruited as participants. All participants were right-handed, had normal or corrected-to-normal vision and were screened for history of psychiatric and neurological disorders. Written informed consents and confidentiality statements were obtained prior to participation from all participants. A total of 22 participants were excluded due to spontaneous expression of doubts concerning the study's objective (5 in the fMRI experiment, 8 in the anodal tDCS group and 9 in the sham tDCS group) or because they reported unawareness of the possibility of cheating in the debriefing (5 in each tDCS group, see supplementary information for details). In addition, data from 3 participants (1 in the fMRI experiment, 1 in the anodal group, 1 in the sham group) were discarded due to termination of the experiment upon their requests. As the final sample, there were 24 subjects in the fMRI experiment (15 females, mean age ± std. = 18.38 ± 0.50 years). For the tDCS experiment, there were 50 subjects in each group (anodal group: 40 females, mean age ± std. = 18.98 ± 0.77 years; sham group: 40 females, mean age ± std. = 18.90 ± 0.79 years). To ensure adequate statistical power, the final sample size was validated with a power analysis using G*Power for each experiment (fMRI: f = 0.25 with 5% statistical significance and 80% power; tDCS: Cohen's d = 0.5 with 5% statistical significance and 80% power). The experimental protocol was approved by the human subject ethics committee in South China Normal University (Approval number: SCNU-PSY-2022-117) and was carried out in accordance with the Declaration of Helsinki.

Cover story and experimental task

We used a meticulously designed cover story to permit opportunities to obtain dishonest gains in an unsupervised scenario and to allow for inconspicuous trial-wise registration of cheating behaviors. A contextualization procedure was

used to enhance their confidence in privacy and to make the cover story more convincing. Participants were told that the current study aimed to explore the influence of privacy and mental efforts on individual hunches and that they would be randomly assigned to either the 'supervised group' or the 'unsupervised group' based on the result displayed by a computer lottery program. However, the program was scripted so that the participants always ended up in the 'unsupervised group'. Participants were then briefed on the task protocol together with a research assistant who pretended to be a naïve participant assigned to the 'supervised' group. For the experimental task, participants were asked to concentrate and make an intuitive guess of the outcome of a computerized random coin-flip via a corresponding keypress. To facilitate mental efforts, each trial was associated with a stake, and participants received either a gain or a loss of the corresponding amount of money based on the correctness of their guess. Importantly, participants were told that while the 'supervised group' participants would perform the task with their guesses recorded and monitored, and that the key press before the coin-flip was intended only to match the response procedure between the two groups, which would not be recorded. Since their guesses were completely private and invisible, they had to recall their guesses at the end of each trial. To minimize their suspicion concerning the task protocol, participants were also told that the 'supervised group' participants had to make a random keypress to keep the response procedure identical between the two groups, even though their guesses had already been recorded (see **Supplementary information** for details). Participants practiced for 12 trials prior to the formal fMRI and the tDCS experiments. At the *post-hoc* debriefing, 10% of the net income from all trials (down to ￥0) was included in addition to a small sum of basic payment for completing the task as the final payment for participation.

fMRI experiment

The event-related fMRI experiment consisted of 6 functional runs of 42

continuous trials. Each run started with a 4.44-s fixation and lasted for 392.88s. There were six possible trial-wise stake amounts (￥0, 0.6, 1.2, 2.4, 4.8, 9.6), with 7 trials for each amount. Each trial started with the presentation of the stake amount for 2s, meanwhile participants guessed the coin-flip outcome (head or tail). Subsequently, an animation of a flipping coin was displayed for 0.3s and the last frame (i.e., the coin flip outcome) remained on the screen for 3.7s, during which participants were required to recall and report their guess (the decision phase). Notably, the flipping coin animation was also scripted such that, for each stake amount, the displayed outcome differed from participant's guess in 5 out of 7 trials (Mismatch trials) yet matched participant's guess in the other trials (Match trials), which was designed for better experimental efficiency meanwhile minimizing participants' suspicion. In the Mismatch trials, participants needed to weigh between honesty and monetary gains, whereas in the Match trials no such consideration was needed. Hence, stake amount in Mismatch trials indexed the magnitude of potential dishonest gain. Finally, participants observed the amount of money that was won or lost based on their recalled guess for 1 s and prepared for the next trial. Trial order was optimized using M-sequence [55].

Whole-brain fMRI data were acquired on a 3T Siemens MAGNETOM Prisma MRI scanner with a 20-channel phase-array head coil at the MRI center of South China Normal University. Functional volumes encompassing 35 axial slices (thickness: 3mm, gap: 0.5mm) were acquired using a multi-band accelerated EPI sequence (TE: 30ms, TR: 2220ms, flip angle: 90°, FOV: 192 ×192 mm^{2}, acquisition matrix: 64×64). Prior to the functional runs, a T1-weighted, high-resolution whole-brain structural scan (1 mm^{3} isotropic voxel) was acquired for each participant in the same session using a 3D-MPRAGE sequence.

Behavioral analysis and computational modeling

In-scanner behavioral data were collected using E-prime 2.0. For each stake amount, we computed the cheating frequency as the ratio of the Mismatch trials in which the recalled guess matched the actual coin flip outcome yet, by design, differed from the registered guess. A variant of Weibull cumulative distribution function was fitted to the cheating frequency data for each participant [23]:

$$y = f(x|k,a,b,c) = (1-k-c)(1-e^{-\left(\frac{x}{a}\right)^{b}}) + c$$

where x and y denote the stake amount and its corresponding cheating frequency, respectively. The parameters k and c give the upper and the lower bounds of cheating frequency, which could be conceptualized as the intrinsic motivation to cheat in the absence of any potential dishonest gains and the degree to which honesty is preserved regardless of potential dishonest gains, respectively. The shape of the function is determined by a and b. Maximum-likelihood (ML) estimate of the model parameters were obtained by maximizing the logarithm of the likelihood function $L(k,a,b,c|\boldsymbol{y})$, which is given as:

$$\log(L) = \sum_{i=1}^{K}\left[\log\left(\frac{n_i!}{m_i!\,(n_i-m_i)!}\right) + m_i\log(y_i) + (n_i-m_i)\log(1-y_i)\right]$$

where n_i, m_i and y_i represent the total number of Mismatch trials, the observed number of cheating trials and the predicted cheating frequency corresponding to the *i*-th stake amount, respectively. K is the number of different stake amounts in our experiment. Two complementary methods were used to assess the goodness of fit of our model. First, we calculated the model deviance D (i.e. the logarithm of the log-likelihood ratio between the full model and the best-fitting model) for each participant's data as:

$$D = 2\sum_{i=1}^{K}\left[m_i\log\left(\frac{p_i}{y_i}\right) + (n_i-m_i)\log\left(\frac{1-p_i}{1-y_i}\right)\right]$$

where p_i represented the observed cheating frequency for the *i*-th stake amount. A Monte-Carlo simulation procedure was used for statistical

assessment of model deviance for each participant's data. Specifically, we generated 1000 cheating frequency data sets using the best-fitting model as the generating function and computed the deviance value based on each generated data set, thereby yielding a distribution of deviance that we would expect from a participant whose behavioral responses conformed to our model prediction. A two-sided 95%-confidence interval was obtained to test for the empirical model deviance. Based on the estimated model parameters, the honesty threshold m_H could then be obtained as the inverse function of the midpoint between the lower and the upper bounds:

$$m_H = f^{-1}(\frac{c+1-k}{2})$$

The level of motivational conflict associated with stake amount x was operationalized as the absolute distance between x and m_H, such that smaller values indicated larger conflict:

$$\varphi(x) = |x - m_H|$$

The stake-cheating relationship was analyzed using a multi-level regression model for repeated measures (measurements nested within participants) implemented in the lme4 package in R [56]. In this model, cheating frequency measured for each stake amount from all participants served as the dependent variable, with stake amount and honesty threshold entered as the level-1 and the level-2 predictors respectively. Random intercept and slope were also included. All predictors were z-transformed prior to estimation of regression coefficients. Effect of each predictor was characterized with its corresponding regression coefficient, and the interaction effect between the two predictors was examined using a simple slope analysis based on the standard pick-a-point approach [57]. A mixed linear model was used to analyze the relationship between motivational conflict and decision time. The operationalized motivational conflict index was entered as the fixed-effect predictor, and the mean decision time for each non-zero stake amount as the dependent variable.

Subjects were modeled as random effects. The motivational conflict-decision time relationship was characterized as the coefficient of the fixed-effect predictor.

fMRI data analysis

The functional imaging data were processed using the SPM 12 toolbox (https://www.fil.ion.ucl.ac.uk/spm/) based on Matlab 2021b (The MathWorks). Raw functional volumes first went through slice-timing correction and head motion correction. The motion corrected functional volumes were co-registered to the anatomical T1 scan and then warped to the standard stereotactic MNI space. At the final preprocessing stage, the normalized images were smoothed with an isotropic Gaussian kernel (5-mm FWHM). A high-pass filter (cut-off period: 128s) was applied to the spatially smoothed data to remove low frequency noise.

A general linear modeling (GLM) approach was used to analyze the neural activation related to monetary information encoding, honesty threshold, and motivational conflict.

In all subject-level GLMs, the regressors were generated by convolving the canonical hemodynamic response function (HRF) with a boxcar function spanning the decision phase of each trial. Match trials with incorrect self-reported guesses and trials with no behavioral responses were modeled together as ineffective trials with an additional nuisance regressor. Head motion parameters estimated during realignment were included as covariates of no interest. GLM 1 was used to localize the brain regions whose activation tracked trial-wise monetary stakes, irrespective of whether honesty-gain trade-offs were present. In this GLM, Match trials with correct self-reported guess and Mismatch trials with behavioral response were modeled with a single regressor, with trial-wise stake amount entered as a parametric modulator. For each participant, the contrast image corresponding to the parametric modulator of

stake amount was computed and entered into a group-level random-effects analysis using a one-sample t-test to identify regions showing stake-related modulation. GLM 2 was used to identify the brain regions whose task-evoked activation covaried with individual honesty threshold. In GLM 2, Match trials with correct guess, Mismatch trials with zero stake, Mismatch trials with non-zero stakes were modeled with separate regressors. At the group-level, the subject-level contrasts corresponding to Mismatch (non-zero) > Mismatch (zero) were regressed against individual honesty thresholds across participants. To identify the brain regions that signaled motivational conflict, we further included a parametric modulator for Mismatch trials with non-zero stakes, which corresponded to the absolute distance between trial-wise stake amount and the participant's honesty threshold. Subject-level parametric-modulator contrasts were submitted to a one-sample *t*-test at the group level to identify regions whose activation decreased with the absolute stake-threshold distance (i.e. increased with motivational conflict). Moreover, to minimize potential confounds from stake-related monetary processing, voxels showing stake-related modulation were excluded from the group-level results based on GLM 2. Specifically, the group-level map based on GLM 1 was liberally thresholded ($p < 0.005$, uncorrected), and was then used as an exclusive mask for the GLM 2 results. All group-level maps were corrected for multiple comparisons using cluster-level FWE correction ($p < 0.05$). Regions of interest (ROI) were defined as the supra-threshold clusters on the group-level maps. To analyze the honesty threshold-mPFC relationship, we computed the mean contrast estimate within the mPFC cluster. To analyze the conflict-SPC relationship, we used GLM 3 to obtain the activation estimates corresponding to each non-zero stake amount for each subject. In GLM 3, Mismatch trials with different stake amounts were modeled with separate regressors. In addition, Match trials with correct guesses were modeled with another regressor. Mean beta estimates corresponding to each non-zero stake amount regressor within the SPC clusters were then computed.

To investigate the functional interplay between mPFC and SPC, we applied dynamic causal modeling (DCM) for fMRI as implemented in SPM12 to the BOLD signal time series. Three representative nodes were defined for monetary information encoding, honesty-threshold representation, and motivational conflict by placing 8-mm-radius spheres around the global maxima identified from the previously reported group-level GLM activation maps (dACC, mPFC, and left SPC, respectively). For each participant, the first eigenvariate of the BOLD signal was extracted from each sphere (adjusted for effects of interest) and entered into DCM. DCM involved three sets of parameters: the extrinsic driving input into one or more regions, the intrinsic effective connectivity among regions and the bilinear parameters that capture the modulatory effect of experimental manipulations on specific connections. Trials with non-zero stakes were modeled as the driving input entering dACC, consistent with dACC serving as the representative node for stake encoding. We specified two modulatory inputs. First, the presence of potential dishonest gains (i.e., Mismatch trials with non-zero stakes) was modeled to modulate the dACC → mPFC connection. Second, the amount of potential dishonest gains (i.e., trial-wise stake amount in Mismatch trials) was modeled as a parametric modulatory input on the mPFC → SPC connection. We constructed six candidate models of the intrinsic and/or modulatory connectivity architecture of the three nodes, and fitted each model to each participant's data. Random-effects Bayesian model selection (hierarchical Bayesian approach) was then used to identify the best-fitting model by comparing exceedance probabilities. Finally, we conducted group-level inference on the modulatory effects of the winning model. For the presence-related modulation, we tested whether the corresponding parameter differed from zero across participants using a one-sample *t*-test. For the amount-related modulation, we computed the predicted mPFC → SPC coupling at the smallest and largest stake levels and then compared these two values using a paired-sample *t*-test.

tDCS experiment

A two (anodal group vs. sham group)-by-six (￥0, 0.6, 1.2, 2.4, 4.8, 9.6) mixed design was employed in the tDCS experiment. Participants were told the same cover story and briefed of task requirements using a procedure identical to that in the fMRI experiment. Electric stimulation was delivered offline using a high-definition transcranial direct current system (HD-tDCS, Soterix Medical, New York) via Ag/AgCl electrode quintuplets (12mm in diameter) filled with conductive gel, which were positioned in a cap (Brain Vision) based on the extended international 10-20 EEG system. The central electrode (F2) was placed on top of the right prefrontal cortex in correspondence with the mPFC ROI (MNI coordinate: x = 10, y = 23, z = 45) identified in the fMRI experiment, while the other four electrodes (AFz, AF4, FCz, FC4) were placed at equal distance surrounding the central node to form a closed circuit. For the anodal group (*N* = 50), tDCS was continuously applied at an intensity of 2mA for 20 minutes [58-60]. To minimize the sensations at stimulation onset, the current was linearly ramped up and ramped down over 30s before and after the 20-min stimulation period, respectively. For the sham group (*N* = 50), the current went through the same upward and downward ramping period to mimic the warming and tingling sensations at the beginning of active stimulation [38] and was then switched off. Immediately after the stimulation, they completed two blocks of 126 trials, with 15 Mismatch trials and 6 Match trials for each stake amount. The experiment lasted approximately 30 minutes, consistent with the time window over which tDCS effects have been reported to persist[37-39]. We applied the same multi-level regression analysis used for the in-scanner behavioral data set to the tDCS dataset, with stake amount specified as the level-1 predictor and stimulation group (anodal vs. sham) as the level-2 predictor, respectively.

Post-hoc debriefing

After the experiments, participants entered a *post-hoc* debriefing session in which they were interviewed individually to evaluate the effectiveness of our cover story. To evaluate participants' confidence in their privacy, they reported whether they had suspected that their behaviors were actually being monitored and whether they were under any kind of explicit or implicit influence during the experiment. To evaluate participants' suspicion of the study's objective, they were also asked whether they had suspected that the study might have other objectives, and elaborated on their thoughts after giving an affirmative answer. Finally, participants were asked whether they were aware of the possibility of cheating for more money and were encouraged to freely express their thoughts regarding the study. In addition, we also administered a safety questionnaire for the tDCS experiment participants, in which they rated their subjective feelings, on a 5-point visual analog scale, over a series of questions. These questions concerned headache, concentration, mood, anxiety, visual perception, fatigue and skin sensations (tingle and itchiness) under the modulating electrodes. Independent two-sample *t* tests were conducted to compare each of the rating scores between the two groups. Participants were also asked to describe any other side effects or discomfort they experienced during the experiment.

Figure & legends

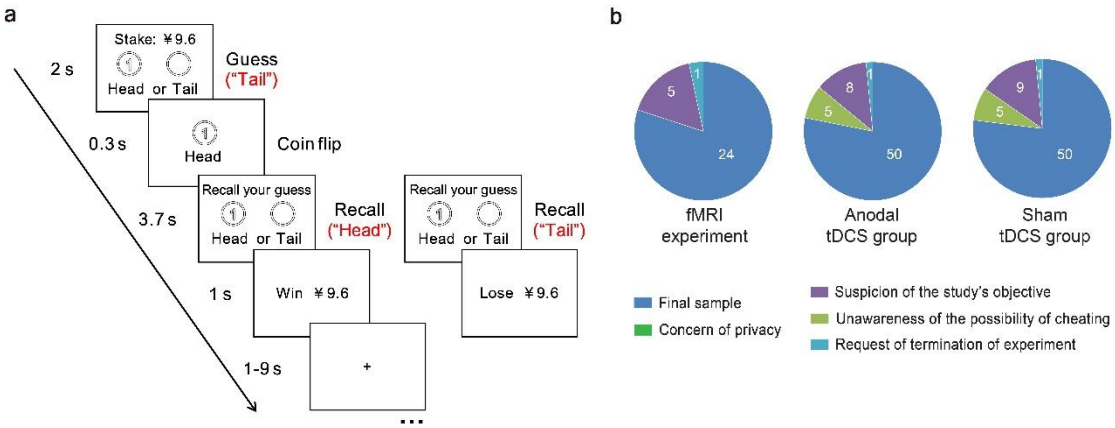


Figure 1. Task protocol and *post-hoc* evaluation of the cover story's

effectiveness. **a)** Procedure of an exemplar Mismatch trial with the participant's response illustrated in red. At the trial beginning, participants were informed of the stake amount and guessed the outcome of an upcoming random coin-flip based on their 'hunch' ('Tail' in this example). The participants' guesses remained completely private and invisible during the experiment. After the coin-flip, the participants were asked to recall their guess, and won or lost the stake based on whether their recalled guess matched the actual coin-flip outcome. In the Match trial, the coin-flip outcome always matched the participant's initial guess. **b)** *Post-hoc* debriefing results of the fMRI experiment (left), the anodal tDCS group (middle) and the sham tDCS group (right).

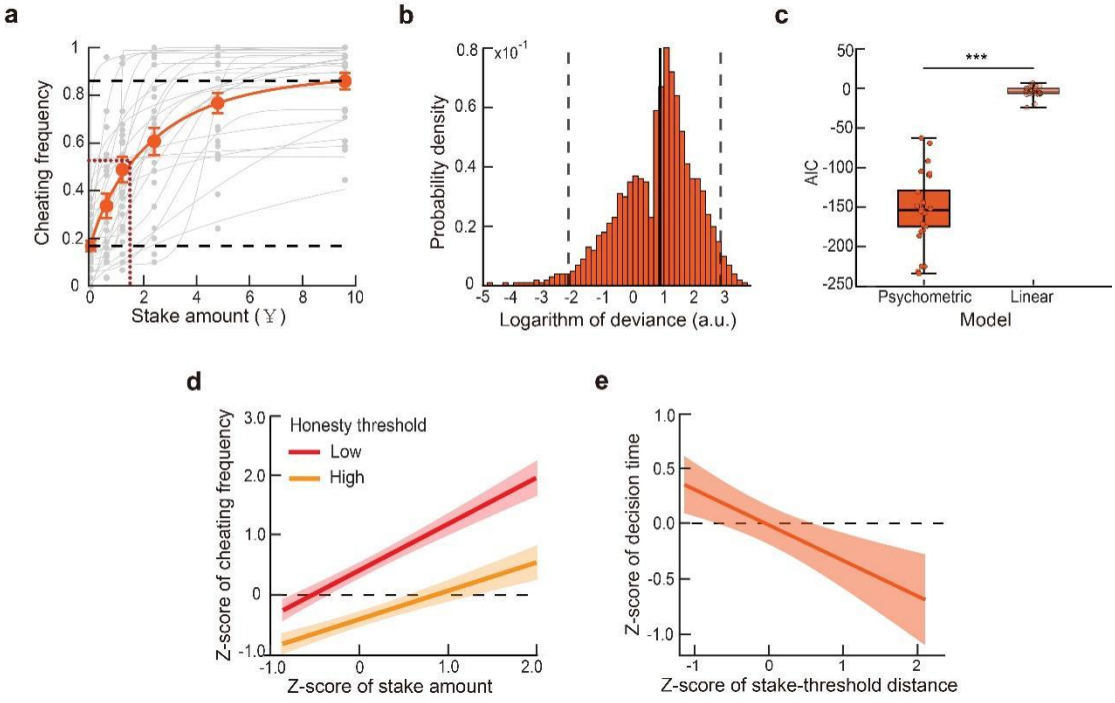


Figure 2. Behavioral results of the fMRI experiment. **a)** Individual (light gray) and group-averaged (colored) cheating frequency as a function of stake amount. Dots denote the measured cheating frequency for the six stake amounts and the curves denote the fitted models. Error bars denote one S.E.M. across all participants. Note that honesty threshold corresponding to the group-averaged model (upper and lower bounds indicated by the horizontal dashed lines) is shown (with dotted lines as visualization aid) for illustration purpose, while in

the actual analyses honesty threshold was identified on an individual basis. **b)** Distribution of deviance of the group-averaged model (in logarithm) obtained using the Monte-Carlo method. Dashed lines indicate the two-sided 95% confidence interval and the solid line indicates the empirical model deviance. The ordinate indicates probability density of the Monte-Carlo distribution. Note that the deviance distribution of the group-averaged model is presented for illustration while the analysis was conducted on an individual basis (see **Figure S1**). **c)** Comparison of AIC values between the psychometric model and the linear model. Boxes denote the bootstrapped 95%-confidence interval of the median AIC values (solid black bars inside the boxes). Errorbars indicate the range of AIC values. Dots denote data from individual subjects. **d)** Influence of honesty threshold on stake-cheating relationship characterized by comparison of simple slopes corresponding to high and the low honesty threshold in the multi-level model. The shaded areas denote 95% confidence interval of the simple slope estimate. Note that both the independent and the dependent variables are z-transformed. **e)** Relationship between decision time and the absolute stake-threshold distance. The shaded area denotes 95% confidence interval of the slope estimate.

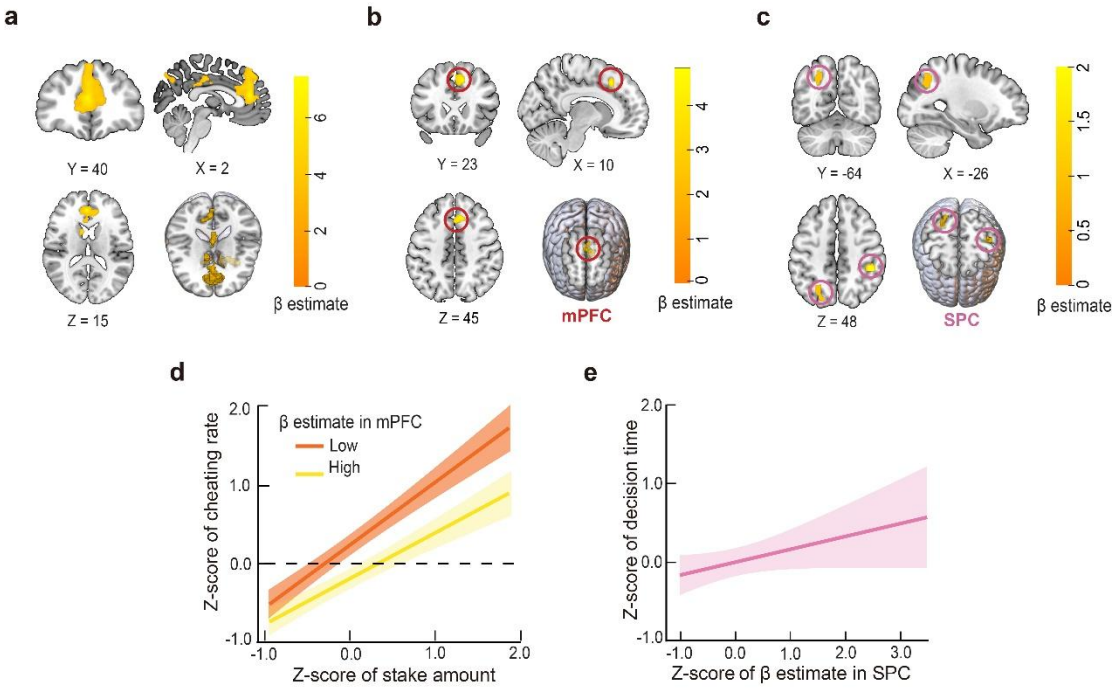


Figure 3. Functional localization results. **a)** Brain regions parametrically modulated by stake amount. **b)** Brain regions positively associated with individual honesty threshold (exclusive mask applied). **c)** Brain regions parametrically modulated by the absolute stake-threshold distance (exclusive mask applied). The slices in **a-c** are displayed according to the MNI coordinates of the global maxima. All maps are corrected for multiple comparisons using whole-brain cluster-level FWE correction ($p < 0.05$). Colored circles highlight the functional clusters (anatomical labels shown inset) selected as region of interest for subsequent analyses in **d-e**. Brain region abbreviations: mPFC: medial prefrontal cortex; SPC: superior parietal cortex. **d)** Influence of mPFC activity on stake-cheating relationship characterized by comparison of simple slopes corresponding to low and high mPFC activation in the multi-level model. Shaded areas denote the 95% confidence interval of the simple slope estimates. **e)** Relationship between decision time and SPC activity across different stake amounts. The solid line indicates the linear fit. The shaded area denotes 95% confidence interval of the slope estimate.

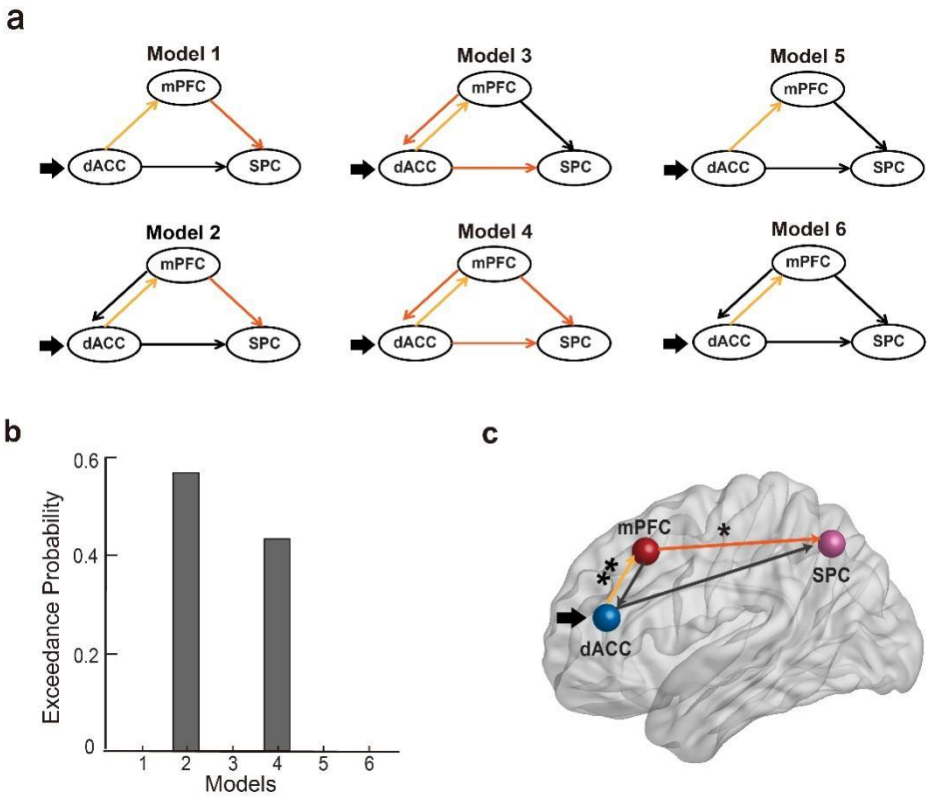
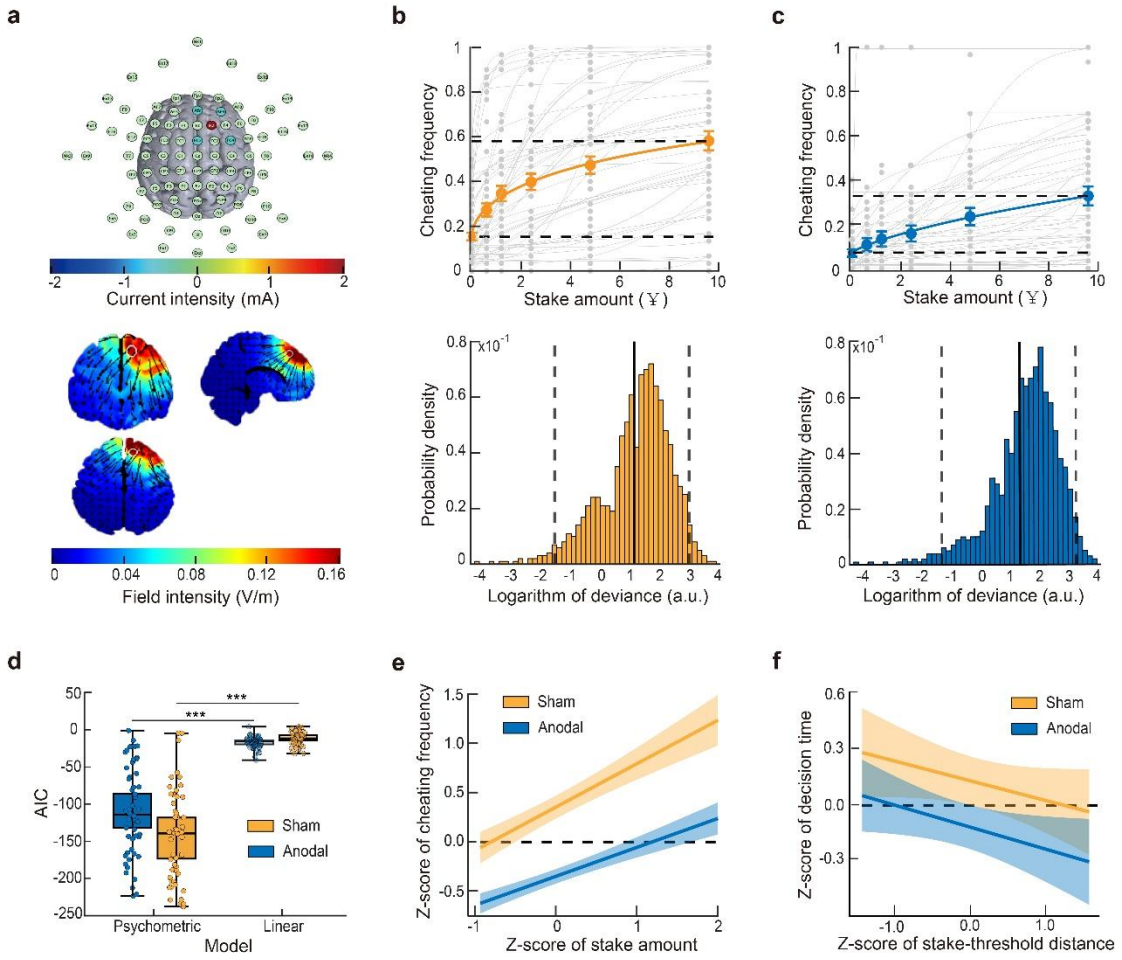


Figure 4. DCM results. **a)** Candidate DCM models. Brain region abbreviations are the same as **Figure 3**. Bold dark arrow denotes the driving input. Yellow arrows denote the effective connectivity modulated by the presence of potential dishonest gains. Orange arrows denote the effective connectivity modulated by the amount of potential dishonest gains. Black arrows denote the unmodulated connectivity. **b)** Comparison of exceedance probabilities of the candidate models. **c)** Group-level inference of modulatory effects, with pattern of effective connectivity corresponding to the winning model rendered on a 3-D brain. *: $p < 0.05$, **: $p < 0.01$.



969 **Figure 5.** tDCS experiment. **a)** Stimulation setup. **Top)** Scheme of stimulation
970 with the 4×1 ring electrode configuration. The central node is shown in red and
971 the 4 equidistant nodes for closed circuit formation are shown in cyan. **Bottom)**
972 Electric field simulation. The direction and length of the arrows indicates the
973 direction and the intensity of the current flows. Behavioral data and model fits
974 of **b)** the sham and **c)** the anodal group. **Top)** Individual data and model fits are
975 shown in light gray, and the group-averaged data and model are shown in
976 colors. Error bars denote one S.E.M. across all participants. **Bottom)**
977 Distribution of group-averaged model deviance (in logarithm) of the sham group
978 and the anodal group. Dashed lines indicate the two-sided 95% confidence
979 interval and the solid lines indicate the empirical model deviance. **d)** Model
980 comparison results of the two groups. Boxes denote the bootstrapped 95%-
981 confidence interval of the median AIC values (solid black bars inside). Errorbars
982 indicate the range of AIC values. Dots denote data from individual subjects. **e)**

Influence of stimulation on stake-cheating relationship characterized by comparison of simple slopes corresponding to the sham group and the anodal group in the multi-level model. The shaded areas denote 95% confidence interval of the simple slope estimate. **f)** Relationship between decision time and the absolute stake-threshold distance in the two groups. The shaded area denotes 95% confidence interval of the simple slope estimate.

Tables

Table 1. Brain regions parametrically modulated by stake amount

Brain region	Cluster size (number of voxels)	MNI Coordinates of cluster peak (mm)			*t*-value
		x	*y*	*z*	
Anterior cingulate cortex	865	2	40	15	7.66
Neucleus accumbens (R)	121	10	6	18	6.94
Putamen (L)	159	-30	10	-10	6.22
Precuneus (R)	83	7	-62	32	5.70
Medial cingulate cortex	73	2	-22	40	5.38

Abbreviations: R: right hemisphere, L: left hemisphere

Table 2. Brain regions positively associated with individual honesty threshold

Brain region	Cluster size (number of voxels)	MNI Coordinates of cluster peak (mm)			*t*-value
		x	*y*	*Z*	
Medial Prefrontal Cortex (R)	136	10	23	45	5.04

Abbreviations: R: right hemisphere, L: left hemisphere

Table 3. Brain regions parametrically modulated by the operationalized motivational conflict index

Brain region	Cluster size (number of	MNI Coordinates of cluster peak (mm)	*t*-value

	voxels)	*x*	*y*	*Z*	
Superior Parietal Cortex (R)	72	44	-37	48	5.31
Superior Parietal Cortex (L)	126	-26	-64	48	4.99

Abbreviations: R: right hemisphere, L: left hemisphere

References

1. Cohn A, Marechal MA, Tannenbaum D, Zund CL. Civic honesty around the globe. *Science* **365**, 70-73 (2019).
2. Yang Q*, et al.* Unraveling controversies over civic honesty measurement: An extended field replication in China. *Proc Natl Acad Sci U S A* **120**, e2213824120 (2023).
3. Shalvi S*, et al.* The science of honesty: A review and research agenda. In: *Advances in Experimental Social Psychology* (ed Gawronski B). Academic Press (2025).
4. Ariely D, Bracha A, Meier S. Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *Am Econ Rev* **99**, 544-555 (2009).
5. Dhar R, Wertenbroch K. Self-Signaling and the Costs and Benefits of Temptation in Consumer Choice. *J Marketing Res* **49**, 15-25 (2012).
6. Gneezy U, Kajackaite A, Sobel J. Lying Aversion and the Size of the Lie. *Am Econ Rev* **108**, 419-453 (2018).
7. Zhu L*, et al.* Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nat Neurosci* **17**, 1319-1321 (2014).
8. Mazar N, Amir O, Ariely D. The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *J Mark Res* **45**, 633-644 (2008).
9. Lockwood PL, van den Bos W, Dreher JC. Moral Learning and Decision-Making Across the Lifespan. *Annu Rev Psychol* **76**, 475-500 (2025).
10. Abe N*, et al.* The neural basis of dishonest decisions that serve to harm or help the target. *Brain Cogn* **90**, 41-49 (2014).
11. Greene JD, Paxton JM. Patterns of neural activity associated with honest and dishonest moral decisions. *Proc Natl Acad Sci U S A* **106**, 12506-12511 (2009).
12. Speer SPH, Smidts A, Boksem MAS. Cognitive control increases honesty in cheaters but cheating in those who are honest. *P Natl Acad Sci USA* **117**, 19080-19091 (2020).
13. Yin LJ, Reuter M, Weber B. Let the man choose what to do: Neural correlates of spontaneous lying and truth-telling. *Brain Cogn* **102**, 13-25 (2016).
14. Hu Y, Hu C, Derrington E, Corgnet B, Qu C, Dreher JC. Neural basis of corruption in power-holders. *Elife* **10**, e63922 (2021).
15. Qu C, Benistant J, Dreher JC. Neurocomputational mechanisms engaged in moral choices and moral learning. *Neurosci Biobehav Rev* **132**, 50-60 (2022).
16. Pang L, Li H, Liu Q, Luo YJ, Mobbs D, Wu H. Resting-state functional connectivity of social brain regions predicts motivated dishonesty. *Neuroimage* **256**, 119253 (2022).
17. Speer SPH, Smidts A, Boksem MAS. Cognitive control and dishonesty. *Trends Cogn Sci* **26**, 796-808 (2022).
18. Yin L, Zhong S, Guo X, Li Z. Functional connectivity between the caudate and medial prefrontal cortex reflects individual honesty variations in adults and children.

Neuroimage **238**, 118268 (2021).
19. Marechal MA, Cohn A, Ugazio G, Ruff CC. Increasing honesty in humans with noninvasive brain stimulation. *Proc Natl Acad Sci U S A* **114**, 4360-4364 (2017).
20. Hu Y*, et al.* Perturbation of Right Dorsolateral Prefrontal Cortex Makes Power Holders Less Resistant to Tempting Bribes. *Psychol Sci* **33**, 412-423 (2022).
21. Thielmann I, Spadaro G, Balliet D. Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychol Bull* **146**, 30-90 (2020).
22. Hilbig BE, Thielmann I. Does everyone have a price? On the role of payoff magnitude for ethical decision making. *Cognition* **163**, 15-25 (2017).
23. Wichmann FA, Hill NJ. The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept Psychophys* **63**, 1293-1313 (2001).
24. Wichmann FA, Hill NJ. The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Percept Psychophys* **63**, 1314-1329 (2001).
25. Goldenberg A*, et al.* Amplification in the evaluation of multiple emotional expressions over time. *Nat Hum Behav* **6**, 1408-1416 (2022).
26. Millgram Y, Nock MK, Bailey DD, Goldenberg A. Knowledge About the Source of Emotion Predicts Emotion-Regulation Attempts, Strategies, and Perceived Emotion-Regulation Success. *Psychol Sci* **34**, 1244-1255 (2023).
27. Schultz W. Multiple reward signals in the brain. *Nat Rev Neurosci* **1**, 199-207 (2000).
28. Haber SN, Knutson B. The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology* **35**, 4-26 (2010).
29. Berridge KC, Robinson TE. Parsing reward. *Trends Neurosci* **26**, 507-513 (2003).
30. Bush G*, et al.* Dorsal anterior cingulate cortex: a role in reward-based decision making. *Proc Natl Acad Sci U S A* **99**, 523-528 (2002).
31. Uddin LQ, Iacoboni M, Lange C, Keenan JP. The self and social cognition: the role of cortical midline structures and mirror neurons. *Trends Cogn Sci* **11**, 153-157 (2007).
32. Gusnard DA, Akbudak E, Shulman GL, Raichle ME. 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).
33. Powell LJ, Macrae CN, Cloutier J, Metcalfe J, Mitchell JP. Dissociable neural substrates for agentic versus conceptual representations of self. *J Cogn Neurosci* **22**, 2186-2197 (2010).
34. Pornpattananangkul N, Zhen S, Yu R. Common and distinct neural correlates of self-serving and prosocial dishonesty. *Hum Brain Mapp* **39**, 3086-3103 (2018).
35. Uddin LQ, Kaplan JT, Molnar-Szakacs I, Zaidel E, Iacoboni M. Self-face recognition activates a frontoparietal "mirror" network in the right hemisphere: an event-related fMRI study. *Neuroimage* **25**, 926-935 (2005).
36. Keenan JP, Gallup GG, Falk D. *The face in the mirror: The search for the origins of consciousness*. HarperCollins Publishers (2003).
37. Shekhawat GS, Stinear CM, Searchfield GD. Transcranial direct current stimulation intensity and duration effects on tinnitus suppression. *Neurorehabil Neural Repair* **27**, 164-172 (2013).
38. Reinhart RM, Cosman JD, Fukuda K, Woodman GF. Using transcranial direct-current stimulation (tDCS) to understand cognitive processing. *Atten Percept Psychophys* **79**,

3-23 (2017).
39. Miler JA, Meron D, Baldwin DS, Garner M. The Effect of Prefrontal Transcranial Direct Current Stimulation on Attention Network Function in Healthy Volunteers. *Neuromodulation* **21**, 355-361 (2018).
40. Nitsche MA, Paulus W. Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans. *Neurology* **57**, 1899-1901 (2001).
41. Jacobson L, Koslowsky M, Lavidor M. tDCS polarity effects in motor and cognitive domains: a meta-analytical review. *Exp Brain Res* **216**, 1-10 (2012).
42. Speer SP, Smidts A, Boksem MAS. Cognitive Control Promotes Either Honesty or Dishonesty Depending on One's Moral Default. *J Neurosci* **41**, 8815-8825 (2021).
43. Thielmann I, Hilbig BE, Zettler I, Moshagen M. On Measuring the Sixth Basic Personality Dimension: A Comparison Between HEXACO Honesty-Humility and Big Six Honesty-Propriety. *Assessment* **24**, 1024-1036 (2017).
44. Tourangeau R, Yan T. Sensitive questions in surveys. *Psychol Bull* **133**, 859-883 (2007).
45. Dunning D, Heath C, Suls JM. Flawed Self-Assessment: Implications for Health, Education, and the Workplace. *Psychol Sci Public Interest* **5**, 69-106 (2004).
46. Paulhus DL. Two-component models of socially desirable responding. *J Pers Soc Psychol* **46**, 598 (1984).
47. Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD. An fMRI investigation of emotional engagement in moral judgment. *Science* **293**, 2105-2108 (2001).
48. Moll J, de Oliveira-Souza R, Bramati IE, Grafman J. Functional networks in emotional moral and nonmoral social judgments. *Neuroimage* **16**, 696-703 (2002).
49. Greene J, Haidt J. How (and where) does moral judgment work? *Trends Cogn Sci* **6**, 517-523 (2002).
50. Qu C, Météreau E, Butera L, Villeval MC, Dreher J-C. Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and social image. *Plos Biol* **17**, e3000283 (2019).
51. Vanhaudenhuyse A*, et al.* Two distinct neuronal networks mediate the awareness of environment and of self. *J Cogn Neurosci* **23**, 570-578 (2011).
52. Qin P, Northoff G. How is our self related to midline regions and the default-mode network? *Neuroimage* **57**, 1221-1233 (2011).
53. Pesenti M, Thioux M, Seron X, De Volder A. Neuroanatomical substrates of arabic number processing, numerical comparison, and simple addition: a PET study. *J Cogn Neurosci* **12**, 461-479 (2000).
54. Simon O, Mangin JF, Cohen L, Le Bihan D, Dehaene S. Topographical layout of hand, eye, calculation, and language-related areas in the human parietal lobe. *Neuron* **33**, 475-487 (2002).
55. Buracas GT, Boynton GM. Efficient design of event-related fMRI experiments using M-sequences. *Neuroimage* **16**, 801-813 (2002).
56. Bates D, Mächler M, Bolker BM, Walker SC. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* **67**, 1-48 (2015).
57. Jaccard J, Turrisi R. *Interaction effects in multiple regression*. Sage (2003).
58. Shen B, Yin Y, Wang J, Zhou X, McClure SM, Li J. High-definition tDCS alters impulsivity in a baseline-dependent manner. *Neuroimage* **143**, 343-352 (2016).

59. Shekhawat GS*, et al.* Intensity, Duration, and Location of High-Definition Transcranial Direct Current Stimulation for Tinnitus Relief. *Neurorehabil Neural Repair* **30**, 349-359 (2016).
60. Gozenman F, Berryhill ME. Working memory capacity differentially influences responses to tDCS and HD-tDCS in a retro-cue task. *Neurosci Lett* **629**, 105-109 (2016).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.docx