JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Neural representation in MPFC reveals hidden selfish motivation in white lies

https://doi.org/10.1523/JNEUROSCI.0088-21.2021

Cite as: J. Neurosci 2021; 10.1523/JNEUROSCI.0088-21.2021

Received: 14 January 2021
Revised: 10 April 2021
Accepted: 16 April 2021

Title**: Neural representation in MPFC reveals hidden selfish motivation in white lies.**

Abbreviated Title: Selfish Motivation in White Lies

JuYoung Kim[1], Hackjin Kim[1*]

[1]Laboratory of Social and Decision Neuroscience

[2]Department of Psychology,

Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea

*Correspondence should be addressed to Hackjin Kim (hackjinkim@korea.ac.kr)

Number of pages: 42

Number of figures: 8

Number words for: Abstract (155), Introduction (644), Discussion (1493)

Conflict of interest: The authors declare no conflicts of interests.

Acknowledgments: This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2018S1A3A2075114)

ABSTRACT

Identifying true motivation for Pareto lies, which are mutually beneficial for both the liar and others, can be challenging because different covert motivations can lead to identical overt behavior. In this study, we adopted a brain-fingerprinting approach, combining both univariate and multivariate analyses to estimate individual measures of selfish motivation in Pareto lies by the degree of multivoxel neural representation in the medial prefrontal cortex (MPFC) for Pareto lies conforms with those for selfish vs. altruistic lies in human participants of either sex. An increase in selfish motivation for Pareto lies was associated with higher mean-level activity in both ventral and rostral MPFC. The former showed an increased pattern similarity to selfish lies and the latter showed a decreased pattern similarity to altruistic lies. Higher ventral MPFC pattern similarity predicted faster response time in Pareto lies. Our findings demonstrated that hidden selfish motivation in white lies can be revealed by neural representation in the MPFC.

SIGNIFICANCE STATEMENT

True motivation for dishonesty serving both self and others cannot be accurately discerned from observed behaviors. Here we showed that fMRI combining both univariate and multivariate analyses can be effectively used to reveal hidden selfish motivation of Pareto lies serving both self and others. The present study suggests that selfish motivation for prosocial dishonesty is encoded primarily by increased activity of the ventromedial and the rostromedial prefrontal cortex, representing intuitive self-serving valuation and strategic switching of motivation depending on beneficiary of dishonesty, respectively.

INTRODUCTION

The consequences of dishonest behavior regarding oneself or others are the key elements that drive dishonesty. Recent studies have reported neural processes associated with prosocial and selfish goals of dishonesty (Yin et al., 2017; Cui et al., 2018). However, less is known about Pareto lies (Erat and Gneezy, 2012), where the results of dishonesty are mutually beneficial for both the liar and others. Two different psychological mechanisms have been proposed to contribute to increasing Pareto lies. The presence of another beneficiary 1) may help justify dishonesty that will benefit oneself, or 2) may trigger genuine care and concern about the benefits others receive (Gino et al., 2013). As Pareto lies are both self-serving and altruistic, recognizing the exact mechanisms engaged from the dishonest behavior alone poses a challenge.

We applied the concept of brain fingerprinting technique (Ahuja and Singh, 2012) to neuroimaging data to gain further evidence for inferring an individual's covert motivation of Pareto lies. In this approach, a target without an explicit label can be classified based on the degree to which the brain response to the target resembles the two known categories. More specifically, contexts in which dishonesty may benefit both self and others may appear as a selfish opportunity to some as they may benefit from dishonesty, whereas the same context may be viewed by others as an altruistic opportunity to benefit others.

Several sub-regions of the medial prefrontal cortex (MPFC) serve a crucial role in moral judgment and generation of dishonest behavior. For example, judgments of the dishonesty of a scenario activate the dorsomedial prefrontal cortex (DMPFC) (Parkinson et al., 2011), spontaneous lying engages the subgenual anterior cingulate cortex (Yin et al., 2016), and the ventromedial prefrontal cortex (VMPFC) is involved in deceiving others (Abe et al., 2007) regardless of whether dishonesty is beneficial to the liar or others

(Pornpattananangkul et al., 2018). Sub-regions of MPFC have been reported to represent self- and other-regarding values differently, where the individual differences in prosociality are expressed as the spatial gradient along the dorsal-to-ventral axis in representing self- and other-regarding values (Sul et al., 2015). Importantly, the rostromedial prefrontal cortex (RMPFC), which includes the pregenual anterior cingulate cortex (Vogt, 2005), is known for computing the values of the outcomes that benefit both self and others (Hutcherson et al., 2015; Sul et al., 2015) and of context-dependent strategic social decisions (Jung et al., 2018; Yoon et al., 2018). It was recently suggested that MPFC subregions are hierarchically organized so that more dorsal regions utilize additional external sensory information from the environment to regulate more ventral subregions that compute intuitive social valuation based on internal bodily signals (Kim, 2020). From this perspective, we predicted that distinctive patterns of activity across MPFC subregions would reflect individual differences in motivation for Pareto lies, given the roles of VMPFC and RMPFC in intuitive social valuation and context-dependent strategic social valuation, respectively.

We aim (1) to identify individuals' primary motivation behind Pareto lies, and (2) to examine neural mechanisms that underlie the processing of immoral opportunities to gain from Pareto lies as opportunities to justify selfish gain, particularly focusing on the differential engagement of the MPFC subregions. To this end, we devised a behavioral task that could measure selfish and altruistic lies as well as Pareto lies. Participants took part in a dot-discrimination task inside the MRI scanner, where they could gain points that would later reduce the length of the stressful task for themselves, another person, or both by being dishonest in each trial (Fig. 1B). We applied both the univariate and multivariate analyses to probe the neural mechanism that underlies the individual difference in the selfish motivation for Pareto lies, as univariate tests may detect the regions mapping the subject-level variability

93 in the selfish motivation, but may not be sensitive enough to reveal the latent sub-features
94 between the conditions within an individual (Davis et al., 2014).
95

MATERIALS AND METHODS

Participants

Forty-three participants (16 females, mean age = 23.79 ± 2.49) were recruited through Korea University students' community website. The following seven participants were excluded from the analyses: three participants for later reporting not to have believed in the experiment cover story, two for misunderstanding the instruction, one for reporting a neuropsychological drug's intake, and one for sleeping during the main task. Behavioral and neuroimaging data of 36 participants were included in the analyses. A power analysis for an rmANOVA testing for within factors suggested that the appropriate sample size to achieve a power of 0.95 with an α of 0.05 and an effect size of .31 was 32. The effect size used in the power analysis was calculated from the partial η^2 of beneficiary × point interaction taken from an independent behavioral pilot study. All participants gave written consent before participation and were compensated with KRW 30,000 (roughly equivalent to USD 30). The study design and the data collection procedures complied with all relevant ethical regulations and were approved by the Korea University Institutional Review Board.

Experimental design and statistical analyses

Experimental Procedure

Participants were given the following overall instruction and a cover story upon arrival. We informed all the participants that the study was about the change in the subjective experience of stressful noise after the depletion of cognitive resources resulting from performing a cognitive task that requires attention. They were to be exposed to an aversive noise for ten minutes and would have to report the stressfulness of the noise after the main task in the MRI scanner. The subject-specific noise level participants would be later exposed to were

121 determined through the noise thresholding procedure to ensure that every subject would
122 experience the same level of evoked stress. They would earn the points in the main cognitive
123 task for themselves or their partner, and the earned points were to be used to reduce the
124 duration of exposure to the stressful noise for the respective beneficiary. The partner they
125 were obtaining points for was another person, unknown to the participant, that would
126 participate in the same experiment immediately after the participant. All participants were
127 told that the same procedure was done for the previous participant, but the amount of points
128 obtained by the previous participant for them was untold. Reduction of the stressful task was
129 used instead of monetary gain as reward for dishonest gain, because controlling for the
130 subjective value of each point across participants in the absence of beneficiary was crucial as
131 our goal was to observe differences in the motivation behind dishonesty for different
132 beneficiaries. The value of each point was manipulated to be similar across participants
133 through the noise thresholding procedure.

134 Following the overall instruction (Fig. 1A), each participant went through the noise
135 thresholding and the dot-screen display time calibration procedures before participating in the
136 main task. Dot-screen display time calibration and the main task was performed inside the
137 scanner. We introduced the dot-screen display time calibration procedure as practice trials.
138 The stressfulness rating, which participants believed they would have to participate in after
139 the main task, did not actually take place.

140 ***Noise thresholding procedure***

141 Participants listened and evaluated a series of brief sounds with differing frequency and
142 volume on a 10-point averseness scale. Participant-specific noise thresholds were determined
143 as the sound each participant evaluated as eight on the 10-point averseness scale. This
144 procedure allowed controlling for the subjective value of points to be obtained during the
145 main task.

Dot-screen display time calibration procedure

Participants performed a simpler version of the dot-discrimination task prior to the main task. They were asked to report the side with more dots. We lengthened the dot screen display time duration when the participant reported the wrong side on the previous trial until each participant could provide correct answers in 10 consecutive trials. The final length of the display time determined by this procedure was used as the dot-screen display time customized for each participant in the main task. We adopted this procedure to ensure that the participants' dishonest decisions in the main task were the intended dishonesty, rather than a perceptual mistake. However, this procedure was introduced as a practice trial, and participants were unaware of the intention behind the procedure.

Dot-discrimination task

The task was introduced to the participants as a visual perception and attention task, and participants were instructed to report the side of the screen with more dots. In each trial, the beneficiary and the points assigned to each side of the screen were shown before the dot-screen appeared. The reward magnitude (i.e., number of points) and the beneficiary of the dishonest decision were experimentally manipulated and varied across trials. We displayed the dot-screen for the individually calibrated length of time which was just long enough for the participant to be aware of the difference in the number of dots between the two sides. Points could only be obtained by being dishonest, that is, by choosing the side with fewer dots, and could benefit the participant (Self), their partner (Other), or both the participant and the partner (Both) (Fig. 1B). The number of points ranged from 0 to 2 points and the points obtained in the five randomly selected trials across conditions were to be used to reduce the duration of the exposure to the stressful noise after the task, and each point would reduce 10 seconds of the total duration. 20 trials existed per each condition, resulting in 180 trials in total. Thus, a single trial consisted of a fixation period (2 ~ 4 s), the beneficiary information

display (0.5 sec) followed by the number of points assigned to each side of the screen (1~3s), dot-screen display for the individually calibrated length of time, question display (until decision), and the result of choice display (0.7 sec).

Behavioral data analyses

The overall effect of point and beneficiary on dishonest decisions was assessed by entering the percentage of wrong choices to a repeated-measures analysis of variance (rmANOVA) with the beneficiary (Self, Other, Both) and point (0, 1, 2) as within-subject factors. As the points could only be obtained by reporting the wrong answer, we expected a higher percentage of dishonest decisions in point 1 and 2 conditions compared to point 0 condition.

We first normalized the RT data within each subject over all trials, and then averaged them separately for dishonest decisions in each condition. For participants who were always honest in certain conditions and whose average RT could not be calculated were excluded from correlation analyses that includes RT data. The correlation between RT data and other indices were obtained using Spearman's rank correlation as the sample size after exclusion resulted in 28, which may be insufficient to use Pearson's correlation. The average normalized RT of each condition were calculated and entered in the rmANOVA for all the 36 participants.

Neuroimaging procedures and analyses

FMRI data acquisition and preprocessing

FMRI data were acquired using a 3.0 T Siemens Magnetom Trio MRI scanner with a 12-channel head matrix coil located at the Korea University Brain Imaging Center. We obtained the T2*-weighted functional images using gradient-echo echo-planar pulse sequences (repeat time (TR) = 2000 ms; echo time (TE) = 30 ms; flip angle (FA) = 90, field of view (FOV) =

240 mm, 80 × 80 matrix; 36 slices; voxel size = 3 mm × 3 mm × 3 mm). The FMRI blood oxygen-level dependent (BOLD) activity was measured over one functional run, lasting about 25 minutes. We acquired the EPI volumes at an oblique angle to the AC-PC line to decrease the impact of susceptibility artifacts in the orbitofrontal cortex. High-resolution T1-weighted (TR = 1900 ms; TE = 2.52 ms; flip angle = 9; 256 × 256 matrix; 1 × 1 × 1 mm in-plane resolution) structural images and diffusion tensor scans (TR = 3000 ms; TE = 70.0 ms; 224 × 224 matrix; voxel size = 2 mm × 2 mm × 2 mm) were also obtained. The stimuli were presented through an MR-compatible liquid-crystal display monitor mounted on a head coil (refresh rate: 85 Hz; display resolution: 800 × 600 pixels; viewing angle: 30 horizontal, 23 vertical).

We preprocessed the data using the SPM12 (Wellcome Department of Imaging Neuroscience, University College of London, London, UK). Images were temporally corrected for interleaved slice acquisition, and then realigned to the first volume to correct for head motion and a mean image was created for each participant. The realigned images were normalized to the standard Montreal Neurological Institute (MNI) EPI template, resampled to 2 × 2 × 2 mm voxels, and spatially smoothed using a Gaussian kernel with an 8 mm full width at half maximum (FWHM).

1st level univariate analyses

A first-level generalized linear model (GLM) was estimated to create contrasts for each beneficiary condition. Onset times for the three beneficiaries (Self, Other, and Both), with the three points (0, 1, and 2 points) information presentation and decisions for each nine condition as well as six head-motion parameters were included as regressors after being convolved with a standard hemodynamic response function. The brain regions reflecting the point by beneficiary interaction effect were identified by first generating three contrast images (i.e., one for each beneficiary condition) by combining Point 1 and 2 conditions and

subtracting point 0 condition at decision onset (e.g., [Point 1+ Point 2] − Point 0 for Self condition), and then entering the contrasts into an rmANOVA. These three contrasts were used in the pattern classification analyses as well. We used these contrasts rather than the contrast of dishonest vs. honest decisions because 1) some participants do not have enough trials of dishonest decision in some conditions, 2) the focus of this research was to distinguish individual motivation and neural mechanisms that underlie the processing of immoral opportunities to gain from Pareto lies.

2nd level univariate analyses

To explore brain regions representing the main effects of beneficiary and point, and the interaction effect between beneficiary and point, three rmANOVAs were conducted. The beneficiary main effect was assessed by constructing first-level contrast images for each beneficiary at decision onset by combining trials overall points for each beneficiary (i.e., Point 0 + Point 1 + Point 2 separately for each of Self, Other, and Both), which were entered into a rmANOAV. In addition, contrasts for each point overall beneficiaries were built and entered into a rmANOVA to examine the main effect of the points. All the statistical maps reported were thresholded at the whole-brain FWE corrected $p < 0.05$ at voxel-level.

Neural signatures of selfish or altruistic motivation for dishonesty: Multivariate analysis

A total variation L1 (TV-L1) pattern classifier (Gramfort et al., 2013) was trained to distinguish between neural patterns associated with the opportunities to lie for Self and Other at the moment of decision. The analysis was performed with Nilearn and nltools library in Python 3 (Abraham et al., 2014). For each beneficiary, representations in the MPFC of the dishonest opportunities were obtained from individual contrasts combining Point 1 and 2 conditions and subtracting point 0 condition at decision onset (e.g., [Point 1+ Point 2] − Point 0 for Self condition). As the primary aim of this analysis was to identify individuals' motivation when dishonest opportunities were given to gain for Both, we contrasted the

conditions in which participants were motivated to lie (i.e., points would be given when lying) with the condition in which participants had no reason to lie (i.e., no point would be given when lying) for each beneficiary. Supporting our rationale for this analysis, the behavioral data showed that participants were induced to lie by the existence, rather than the amount, of available point to be earned. The classifier was first trained on the MPFC activity pattern for Self and Other beneficiary conditions to distinguish between neural patterns associated with the opportunities to lie for Self and Other. Conducting classification using moderately smoothed data is thought to be effective (Op de Beeck, 2010; Hendriks et al., 2017) especially when the objective of the classification is to generalize across subjects (Chang et al., 2015; Weaverdyck et al., 2020). The MPFC binary mask was taken from a meta-analysis segregating MPFC into sub-regions based on the each region's functional co-activation maps (De La Vega et al., 2016). Of the nine MPFC sub-regions, we excluded the supplementary motor area (SMA) and pre-SMA from the mask as activity related to movements or movement control was not considered in this study. Eight-fold nested cross validation was applied where 10 contrast images out of 72 were held out as test data, and the remaining 62 images were used as the training data at each fold. The best performing weights were selected at each fold, and the final classifier weight map was constructed by taking the average of the weights of the overall folds. The MPFC activity pattern of the individual first-level contrast maps of Both condition were entered into the trained classifier to predict the class of each individual's MPFC activity pattern of Both condition (Fig. 3A).

The individual measure of selfish motivation in Pareto lies was defined as how certain each individual's MPFC activity pattern during Both condition was classified as Self. As such, the signed distance of individual Both contrast to the hyperplane separating Self and Other was calculated and used as the self-class confidence scores (SCCS), where a higher score translates into higher certainty of being classified into Self. Computationally, this score

was calculated by taking the dot product of individual Both contrast and the classifier weight map and adding the intercept term.

Second-level regression and correlation analyses with the SCCS

Multiple regression analyses were performed to explore the neural mechanisms behind selfish motivation in each beneficiary's opportunities. In these analyses, the SCCS were regressed on the contrast maps of Self, Other, and Both conditions separately.

Representational similarity analyses

The VMPFC, RMPFC, and precuneus masks were generated from the result of the whole-brain FWE corrected multiple regression analysis of the SCCS with Both contrasts (VMPFC cluster peak: x = -2, y = 46, z = -8; RMPFC cluster peak: x = 8, y = 34, z = 14; precuneus cluster peak: x = 10, y = -60, z = 34). We extracted the neural activity of Self and Both conditions in the ROIs for each participant from the Self, Other, and Both contrasts used in the univariate analyses. We also calculated the pattern similarity as the Kendall's Tau (Popal et al., 2019) between the neural activity patterns in each ROI of Self and Both conditions, and those of Other and Both conditions for each participant. Then, the calculated pattern similarities were correlated with the SCCS.

RESULTS

Behavioral results

We first tested whether participants were more likely to report incorrectly when points were available, as this suggests dishonesty, and whether such dishonesty is modulated by the beneficiary of the points. We carried out a two-way repeated-measures ANOVA (rmANOVA) to assess the effect of points and beneficiary on the participants' decisions to be dishonest. A significant main effect of point ($F(2,70) = 26.971$, $p < .001$) was revealed with a significant linear trend ($F(1,35) = 28.380$, $p < .001$; Fig. 1C) as expected. This suggests that participants were more dishonest as more points were available. The main effect of the beneficiary was also significant ($F(2,70) = 5.078$, $p = .009$), and behavioral patterns indicated that the participants were generally more dishonest when points were available for Self or Both than for Other. Beneficiary by point interaction was significant ($F(4, 140) = 3.075$, $p = .018$), implying that each point had a different impact on dishonest decisions depending on the beneficiary. For each pair of the beneficiaries, we ran a two-by-three rmANOVA with beneficiary and point as factors to investigate the cause of the interaction. The analyses revealed that the beneficiary and point interaction was significant for Self and Other ($F(2, 70) = 4.894$, $p = 0.010$), and Other and Both conditions ($F(2, 70) = 3.722$, $p = 0.029$), but not for Self and Both conditions ($F(2, 70) = 0.278$, $p = 0.758$). We tested for the difference of 2-way interaction terms between pairs of conditions. The interaction of beneficiary and point was calculated as (P1 + P2) / 2 − P0 for each beneficiary, and the difference of the interaction term between each pair of beneficiaries were entered in paired t-tests. The analyses revealed that 2-way interaction of points and beneficiary of Self and Other condition pair was significantly different from the 2-way interaction of Self and Both condition pair ($t(36) = -$

2.514, p = .017), and the 2-way interaction of Other and Both condition pair was significantly different from Self and Both condition pair (t(36) = 2.854, p = .007). This indicates a selectively lower dishonesty rate for Other as opposed to Self and Both conditions. The main effect of the point and beneficiary, and the interaction of the two were not significant for the response time (RT) data, but we observed a significant negative correlation between the ratio of dishonest decisions in Both condition and the RT of dishonest choices in Both condition (Fig. 7C, Spearman's rho(28) = -.561, p = .002, two-sided). This suggests the individuals who were faster in dishonest decisions for Both were more prone to be dishonest in Both condition.

Neuroimaging results

Univariate analysis result

We first investigated how opportunities to gain from dishonesty for different beneficiaries and different amounts of points are represented in the brain. A first-level generalized linear model (GLM1) was built, including onset times for three beneficiaries (Self, Other, and Both), three points (0, 1, and 2 points) information presentation, and decisions for every nine combinations of beneficiaries and points, which were all convolved with a standard hemodynamic response function. The model also included six motion parameters as nuisance regressors. We created the first-level contrasts for each beneficiary (e.g., Point 0 + Point 1 + Point 2 for Self trials), and each point (e.g., Self + Other + Both for Point 0 trials) to examine brain regions showing the difference in the activation at the time of decision based on the beneficiaries and points, and then entered them into two separate second-level rmANOVAs to assess the main effects of beneficiary and points. The analyses revealed a unique RMPFC response for each beneficiary (Fig. 2A, x = 0, y = 40, z = 20; whole-brain FWE corrected at voxel-level p < .05 unless stated otherwise), showing the highest activity during Self, and lowest activity during Both conditions. Furthermore, a larger RMPFC cluster extending into

the DMPFC was revealed to show differences in the activity to the different amounts of points (Fig. 2B, x = 0, y = 40, z = 20), showing higher activity as the available points increased. We assessed the interaction effect between the point and beneficiary using the contrasts constructed by combining conditions where points were available and subtracting the condition where no point was available (i.e., [Point 1 and 2] - Point 0) for each beneficiary for each participant and entering the contrasts into a one-way rmANOVA. The interaction between point and beneficiary was also revealed in a large cluster in the posterodorsal MPFC (x = 0, y = 28, z = 42).

Neural signatures of selfish or altruistic motivation for dishonesty: Univariate analysis

We first conducted a second-level t-test on Self vs Other contrast and Other vs Self contrast to identify the distinctive neural features related to selfish or altruistic motivation for dishonesty. No voxels survived the correction in both contrasts, which confirms our prediction that a univariate analysis may not be sensitive enough for detecting subtle differences in neural representation between selfish and altruistic motivation for lying.

Neural signatures of selfish or altruistic motivation for dishonesty: Multivariate analysis

For a further differentiation of the neural signatures of selfish or altruistic motivation for dishonesty in Both condition, we trained a pattern classifier (For more detailed information, see Methods section 'Multivariate classification of the neural representation of motivations for Pareto lies'**)** to differentiate neural patterns in the MPFC associated with the opportunities to lie for Self and Other. We used the trained classifier to classify individuals' neural patterns for Both conditions to estimate one's covert motivation underlying moral decisions in situations where dishonesty would benefit both Self and Other (Fig. 3A). The classifier was trained across, rather than within, participants to ensure its generalizability. The final classifier showed 98.61% accuracy in distinguishing Self vs Other contrast images. The classification results showed that Both was classified as Self in 17 out of 36 participants and

as Other in the remaining 19 participants. The percentage of Pareto lies would not differ between the two groups ($t(34) = .664$, $p = .511$), consistent with our hypothesis.

Neural evidence for selfish motivation in Pareto lies

Next, we identified neural regions related to the degree of selfish motivation in Pareto lies, which was defined as the self-class confidence scores (SCCS). We calculated the SCCS by taking the signed distance of individuals' Both contrast to the hyperplane separating Self and Other. The SCCS ranged from -2.06 to 2.11 with the mean value of 0.03 and SD of 0.90. The absolute value of the score indicates the certainty of the sample being classified into Self (positive sign) or into Other (negative sign). Thus, the SCCS is assumed to indicate individual differences in the degree of selfish motivation when encountering opportunities to gain from dishonesty for both Self and Other. Individuals' SCCS were then regressed on the contrast map of the Both condition to identify the neural regions uniquely associated with the opportunity for Pareto lies as a function of the degree of selfish motivation. This analysis revealed that the activities in RMPFC ($x = 8$, $y = 34$, $z = 14$; Fig. 4A, C), VMPFC ($x = -2$, $y = 46$, $z = -8$; Fig. 4B, D), and precuneus ($x = 10$, $y = -60$, $z = 34$) positively correlated with the SCCS.

Unlike standard univariate tests, MVPA is now known to be insensitive to inter-subject variability in mean activation across voxels within a region of interest (Davis et al., 2014). Accordingly, we used a representational similarity analysis (RSA) to examine whether multi-voxel patterns in each of the three neural regions associated with the SCCS uniquely encode neural evidence for selfish motivation in Pareto lies. As expected, our analysis revealed that the SCCS correlates *positively* with the degree of similarity of the VMPFC activity pattern between Self and Both conditions (Pearson's $r(36) = .364$, $p = .029$, two-sided; Fig. 5B), but not between Other and Both conditions (Pearson's $r(36) = .123$, $p = .475$, two-sided; Fig. 5D). In addition, the SCCS correlates *negatively* with the degree of similarity

of the RMPFC activity pattern between Other and Both conditions (Pearson's $r(36) = -.401$, $p = .013$, two-sided; Fig. 5C), but not between Self and Both conditions (Pearson's $r(36) = -.071$, $p = .681$, two-sided; Fig. 5A). Tests for differences in dependent correlations showed that the correlation coefficients of Self-Both similarity and Other-Both similarity in RMPFC cluster are significantly different ($z = 1.651$, $p = .049$, one-tailed; Fig. 5E), and the correlation coefficients of Self-Both similarity and Other-Both similarity in VMPFC cluster are marginally different ($z = 1.567$, $p = .057$, one-tailed; Fig. 5F). In the precuneus cluster, the SCCS showed no correlation with the degree of pattern similarity between Self and Both (Pearson's $r(36) = .293$, $p = .082$) nor between Other and Both (Pearson's $r(36) = .306$, $p = .069$).

Behavioral evidence for selfish motivation in Pareto lies

We examined whether this neural evidence for selfish motivation in Pareto lies can be validated by behavioral evidence for Pareto lies. Specifically, we examined the difference between altruistic and Pareto lies as measured by the difference in the proportion of dishonesty and RT between Other and Both conditions. First, as for the VMPFC cluster, the degree of similarity between Self and Both conditions in the activity pattern does not correlate either with the proportion of dishonest choices in Both condition (Pearson's $r(36) = -.066$, $p = .699$) or with the difference in the proportion of dishonest choices between Other vs. Both conditions (Pearson's $r(36) = .106$, $p = .373$, two-sided). However, the same indices show a significant *negative* correlation with the RT of being dishonest in Both condition (Spearman's $\text{rho}(28) = -.492$, $p = .006$, Fig. 7B) and also a significant *positive* correlation with RT differences between Other vs. Both conditions when being dishonest (Spearman's $\text{rho}(28) = .463$, $p = .013$, two-sided; Fig. 7A). These findings suggest that those with a high degree of selfish motivation in lying for Both engage qualitatively different processes subserving altruistic and Pareto lies, which appears mainly due to their faster intuitive

responses in Pareto lies. As for the RMPFC and the precuneus clusters, no significant correlation was found between the representational similarity indices and either Other-Both differences in dishonest decisions (RMPFC: Pearson's $r(36) = .136$, $p = .426$; precuneus Pearson's $r(36) = -.009$, $p = .957$) or those in RT (RMPFC: Spearman's rho(28) = -.016, $p = .934$; precuneus: Spearman's rho(28) = .064, $p = .754$).

Comparing between selfish and altruistic motivations for dishonesty associated with selfish motivation in Pareto lies.

We also examined whether and how selfish motivation in Pareto lies is differentially associated with the neural representations in self- and other-benefiting dishonest opportunities. To achieve this, we regressed the SCCS on the contrast map of the Self and Other conditions separating them into two multiple regression analyses. During Self condition, the activities in the VMPFC ($x = 8$, $y = 44$, $z = -10$) and the ventral striatum (VS: $x = -16$, $y = 8$, $z = -8$) showed significant positive correlations with the SCCS (Fig. 6B). This suggests that as individuals consider opportunities for Both to be closer to opportunities for Self, self-benefiting dishonest opportunities engaged VMPFC and ventral striatum to a larger extent. During the Other condition, a significant positive correlation was observed between individual SCCS and the activities in VMPFC ($x = -6$, $y = 48$, $z = -6$), and VS ($x = -18$, $y = 8$, $z = -2$), similar to the observation made for the Self condition. However, the RMPFC ($x = 6$, $y = 52$, $z = 16$) and left anterior insula (AI: $x = -26$, $y = 20$, $z = -12$) additionally showed significant positive correlations with the SCCS during Other condition (Fig. 6C). These findings indicate that other-benefiting dishonesty additionally engages RMPFC and AI among those Pareto lies that appeared to be primarily driven by selfish motivation.

DISCUSSION

This study proposed to infer individuals' covert primary motivations behind dishonesty based on neuroimaging data by adopting the brain-fingerprinting approach combined with machine-learning. As expected, the exhibited dishonest decisions that profit both the liar and others were identical regardless of the underlying motivation to benefit both. The individual measure of selfish motivation in Pareto white lies was estimated by the degree to which the multivoxel neural representation in the MPFC during Both condition matches that during Self vs. Other condition. The same measures showed positive correlations with the mean level of activity in the VMPFC and the RMPFC during Both condition. Further representational similarity analyses demonstrated that higher selfish motivation in Pareto white lies can be characterized specifically by increased pattern matching between the Both and Self condition in the VMPFC, and decreased pattern matching between the Both and Other condition in the RMPFC. In addition, these neural findings were also mirrored by the behavioral data such that a higher degree of selfish motivation in Pareto lies, as measured by the increased pattern similarity between Self and Both condition in the VMPFC, was associated with faster response times in Pareto vs. altruistic lies, indicating qualitatively different processes subserving altruistic and Pareto lies. In summary, these findings suggest that hidden selfish motivation in white lies can be revealed by neural representation in the MPFC, and increased recruitment, as well as distinctive multivoxel neural patterns, of the VMPFC and the RMPFC characterize selfish motivation in Pareto lies.

Our *a priori* goal of this study was to identify the neural signatures of selfish motivation for Pareto lies. The higher the degree to which multivoxel neural representation in the MPFC during Both condition matches that of Self than Other condition, the larger the mean activity observed in the VMPFC and VS when encountering opportunities for selfish

gain. Moreover, the degree of pattern similarity in the VMPFC between Self and Both condition predicted faster response times for Pareto lies, the possible indicator of impulsive motivation for earning points by lying. Given the well-known functions of VMPFC in processing reward-predicting information (Knutson et al., 2000; O'Doherty et al., 2001; Kim et al., 2011) and intuitive valuation for decision-making (Shenhav and Greene, 2010; Tricomi et al., 2010; Buckholtz and Marois, 2012; Crockett, 2013; Janowski et al., 2013; Sul et al., 2015; Zaki and Cikara, 2015; Jung et al., 2018), these findings suggest that the increased mean activity in the VMPFC, as well as its specific multivoxel representational pattern that is shared between Self and Both conditions, are the core neural evidence and signatures of selfish motivation in Pareto white lies.

Unlike the VMPFC where the mean activity was positively correlated with the SCCS in all three conditions, the higher activity in the RMPFC and AI was positively associated with the SCCS when encountering opportunities for altruistic lies and Pareto white lies, but not for selfish lies. It was recently suggested that the MPFC can be hierarchically organized such that the RMPFC utilizes additional external sensory information from the environment to predict and prevent conflicts occurring in VMPFC tuned to internal bodily signals (Kim, 2020). Consistent with this idea, whereas VMPFC is involved in the internalized/intuitive social valuation, RMPFC contributes to the arbitration between internal and external valuation, playing a key role in context-dependent strategic valuation for social decision-making (Tusche et al., 2016; Jung et al., 2018; Yoon et al., 2018; Cutler and Campbell-Meiklejohn, 2019; Fukuda et al., 2019) including sophisticated and socially appropriate expression of self-protective behavior (Kumaran et al., 2016; Will et al., 2017; Yoon et al., 2018) and socially desirable behavior under social observation (Izuma et al., 2010; Jung et al., 2018; Yoon et al., 2021). Based on these theoretical and empirical studies, we can infer that those with higher selfish motivation in Pareto white lies can be characterized by increased

intuitive/impulsive motivation subserved by the VMPFC and VS when considering dishonesty for Self condition, and also by an increased strategic regulation of such intuitive motivation subserved by the RMPFC and the AI when considering dishonesty for Other and Both.

We also ran RSA on the two regions of interest found in the univariate analysis and demonstrated that the representational *similarity* between Self and Both in the VMPFC and the representational *dissimilarity* between Other and Both in the RMPFC among those with higher SCCS indicating increased selfish motivation of Pareto white lies. Combining these findings with the univariate analysis results, we present the following two arguments. First, the RMPFC clusters showing increased mean activity in both Other and Both conditions may involve distinct neuronal ensembles, each serving different functions. Second, the distinct neuronal ensemble in RMPFC engaged in Both condition, but not in Other condition, may increase the representational *similarity* between Self and Both in the VMPFC cluster (Fig. 8). Those with higher degrees of Self-Both similarity in the VMPFC showed faster response time in Both condition without observable difference in the proportion of dishonesty. This is consistent with the previous findings showing that neural activity related to dishonesty goes in parallel with RT, but not with dishonest behavior (Abe et al., 2018). Increased selfish motivation in Pareto lies likely minimizes conflicts, caused by multiple competing motivations when considering opportunities to lie for both oneself and others. However, those with a greater similarity between Self and Both conditions in the VMPFC activity pattern showed slower response time in being dishonest in Other vs Both condition, potential evidence for qualitatively different mental processes engaged for altruistic and Pareto lies among those with a higher degree of selfish motivation in Pareto lies.

In the univariate analyses, a more posterior cluster in the RMPFC, close to the pregenual anterior cingulate cortex (Vogt, 2005), showed the highest activity when the

beneficiary was Self, and the lowest when the beneficiary was Both. The same region was also more active when more points were available. This activity may not be related to the increased motivation for dishonesty because participants lied more for Both than for Other even to the level of Self, which is opposite to the pattern of neural activity in this region across conditions. This observation led to a more plausible speculation that the activity in this region reflects a conflict between the urge to gain points and the guilt resulting from dishonesty, which is in line with the previous research showing increased ACC activity associated with moral conflict or guilt (Fourie et al., 2014; Abe et al., 2018). The fact that this region showed the lowest activity in Both condition suggests that people experience the least moral conflict when dishonesty can benefit both the liar and another person. In addition, the activity in this region was also stronger among those with higher SCCS, possibly reflecting an increased moral conflict or guilt due to higher selfish motivation for Pareto lies.

We found no evidence for neural signatures of altruistic motivation for Pareto lies because there was no cluster in the brain showing a negative correlation with the SCCS even at a lenient threshold ($p < 0.005$ uncorrected). It has been established that the magnitude of the BOLD response is sensitive to change in excitation-inhibition balance in the cortical microcircuits involving the pyramidal projection neurons interacting with local GABAergic interneurons, which may reflect mismatch or prediction error-related feedback signals (Logothetis, 2008). Given this, larger negative SCCS, or higher Other-classification confidence score may not necessarily involve significant increase in excitation-inhibition balance, because the multivoxel representation analysis can be immune to such a change in excitation-inhibition balance (Logothetis, 2008).

This study provides a novel methodological approach combining the potential benefits of univariate and multivariate analyses. Despite its superior sensitivity to detecting subtle differences in neural representation among different psychological states, MVPA has

not been considered appropriate for identifying the exact neural mechanisms leading to the psychological state at question (Kohoutová et al., 2020), and insensitive to inter-subject variability in mean activation across voxels within a region of interest, which can be better captured by a conventional univariate analysis (Davis et al., 2014). Consistent with the dissociation between univariate and multivariate analyses, multivariate patterns showed a higher similarity of Both to Self vs. Other, whereas univariate patterns showed the opposite, that is, the higher similarity of Both to Other vs. Self, with the RMPFC clusters additionally recruited in Both and Other. This study demonstrated that a univariate analysis can be combined with MVPA to effectively locate the neural regions where the neural representations contributed maximally to the global pattern classification.

In conclusion, this study demonstrates that fMRI can be used to infer hidden selfish motivation in Pareto white lies by adopting the brain fingerprinting approach combining both univariate and multivariate analyses. This technique allowed us to estimate individual differences in motivation for Pareto lies, based on distinctive patterns of activity across functionally dissociable MPFC subregions, including VMPFC and RMPFC. We believe that this study will provide a novel and powerful research method and theoretical contributions to the current efforts of understanding complex motivations underlying moral behaviors.

553 **References**

554

555 Abe N, Greene JD, Kiehl KA (2018) Reduced engagement of the anterior cingulate cortex in
556 the dishonest decision-making of incarcerated psychopaths. Soc Cogn Affect Neurosci
557 13:797–807 Available at: https://academic.oup.com/scan/article/13/8/797/5048611
558 [Accessed August 21, 2020].

559 Abe N, Suzuki M, Mori E, Itoh M, Fujii T (2007) Deceiving others: Distinct neural responses
560 of the prefrontal cortex and amygdala in simple fabrication and deception with social
561 interactions. J Cogn Neurosci 19:287–295.

562 Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A,
563 Thirion B, Varoquaux G (2014) Machine learning for neuroimaging with scikit-learn.
564 Front Neuroinform 8 Available at:
565 http://journal.frontiersin.org/article/10.3389/fninf.2014.00014/abstract [Accessed
566 January 21, 2020].

567 Ahuja D, Singh B (2012) Brain fingerprinting. J Eng Technol Res 4:98–103.

568 Buckholtz JW, Marois R (2012) The roots of modern justice: Cognitive and neural
569 foundations of social norms and their enforcement. Nat Neurosci 15:655–661 Available
570 at: https://www.nature.com/articles/nn.3087 [Accessed August 3, 2020].

571 Chang LJ, Gianaros PJ, Manuck SB, Krishnan A, Wager TD (2015) A Sensitive and Specific
572 Neural Signature for Picture-Induced Negative Affect Adolphs R, ed. PLOS Biol
573 13:e1002180 Available at: https://dx.plos.org/10.1371/journal.pbio.1002180 [Accessed
574 March 31, 2021].

575 Crockett MJ (2013) Models of morality. Trends Cogn Sci 17:363–366.

576 Cui F, Wu S, Wu H, Wang C, Jiao C, Luo Y (2018) Altruistic and self-serving goals
577 modulate behavioral and neural responses in deception. Soc Cogn Affect Neurosci
578 13:63–71.

579 Cutler J, Campbell-Meiklejohn D (2019) A comparative fMRI meta-analysis of altruistic and
580 strategic decisions to give. Neuroimage 184:227–241 Available at:
581 https://doi.org/10.1016/j.neuroimage.2018.09.009.

582 Davis T, LaRocque KF, Mumford JA, Norman KA, Wagner AD, Poldrack RA (2014) What
583 do differences between multi-voxel and univariate analysis mean? How subject-, voxel-,
584 and trial-level variance impact fMRI analysis. Neuroimage 97:271–283 Available at:
585 https://linkinghub.elsevier.com/retrieve/pii/S1053811914003061 [Accessed August 3,
586 2020].

587 De La Vega A, Chang LJ, Banich MT, Wager TD, Yarkoni T (2016) Large-scale meta-
588 analysis of human medial frontal cortex reveals tripartite functional organization. J
589 Neurosci 36:6553–6562.

590 Erat S, Gneezy U (2012) White lies. Manage Sci 58:723–733.

591 Fourie MM, Thomas KGF, Amodio DM, Warton CMR, Meintjes EM (2014) Neural
592 correlates of experienced moral emotion: An fMRI investigation of emotion in response
593 to prejudice feedback. Soc Neurosci 9:203–218 Available at:
594 http://www.tandfonline.com/doi/abs/10.1080/17470919.2013.878750 [Accessed July 13,
595 2020].

596 Fukuda H, Ma N, Suzuki S, Harasawa N, Ueno K, Gardner JL, Ichinohe N, Haruno M,
597 Cheng K, Nakahara H (2019) Computing social value conversion in the human brain. J
598 Neurosci 39:5153–5172 Available at: https://doi.org/10.1523/JNEUROSCI.3117-

18.2019 [Accessed August 3, 2020].

Gino F, Ayal S, Ariely D (2013) Self-serving altruism? The lure of unethical actions that benefit others. J Econ Behav Organ 93:285–292.

Gramfort A, Thirion B, Varoquaux G (2013) Identifying predictive regions from fMRI with TV-L1 prior. In: Proceedings - 2013 3rd International Workshop on Pattern Recognition in Neuroimaging, PRNI 2013, pp 17–20.

Hendriks MHA, Daniels N, Pegado F, de Beeck HPO (2017) The effect of spatial smoothing on representational similarity in a simple motor paradigm. Front Neurol 8:29 Available at: /pmc/articles/PMC5446978/ [Accessed March 29, 2021].

Hutcherson CA, Bushong B, Rangel A (2015) A Neurocomputational Model of Altruistic Choice and Its Implications. Neuron 87:451–462 Available at: http://dx.doi.org/10.1016/j.neuron.2015.06.031.

Izuma K, Saito DN, Sadato N (2010) The roles of the medial prefrontal cortex and striatum in reputation processing. Soc Neurosci 5:133–147 Available at: http://www.tandfonline.com/doi/abs/10.1080/17470910903202559 [Accessed August 10, 2020].

Janowski V, Camerer C, Rangel A (2013) Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. Soc Cogn Affect Neurosci 8:201–208 Available at: https://academic.oup.com/scan/article/8/2/201/1624534 [Accessed August 3, 2020].

Jung D, Sul S, Lee M, Kim H (2018) Social Observation Increases Functional Segregation between MPFC Subregions Predicting Prosocial Consumer Decisions. Sci Rep:1–13 Available at: http://dx.doi.org/10.1038/s41598-018-21449-z.

622 Kim H (2020) Stability or Plasticity? – A Hierarchical Allostatic Regulation Model of Medial
623 Prefrontal Cortex Function for Social Valuation. Front Neurosci 14:1–16 Available at:
624 https://www.frontiersin.org/articles/10.3389/fnins.2020.00281?utm_source=researcher_
625 app&utm_medium=referral&utm_campaign=RESR_MRKT_Researcher_inbound.

626 Kim H, Shimojo S, O'Doherty JP (2011) Overlapping responses for the expectation of juice
627 and money rewards in human ventromedial prefrontal cortex. Cereb Cortex 21:769–776
628 Available at: https://academic.oup.com/cercor/article/21/4/769/286163 [Accessed
629 September 11, 2020].

630 Knutson B, Westdorp A, Kaiser E, Hommer D (2000) FMRI visualization of brain activity
631 during a monetary incentive delay task. Neuroimage 12:20–27.

632 Kohoutová L, Heo J, Cha S, Lee S, Moon T, Wager TD, Woo CW (2020) Toward a unified
633 framework for interpreting machine-learning models in neuroimaging. Nat Protoc
634 15:1399–1435 Available at: https://doi.org/10.1038/s41596-019-0289-5 [Accessed
635 August 17, 2020].

636 Kumaran D, Banino A, Blundell C, Hassabis D, Dayan P (2016) Computations Underlying
637 Social Hierarchy Learning: Distinct Neural Mechanisms for Updating and Representing
638 Self-Relevant Information. Neuron 92:1135–1147.

639 Logothetis NK (2008) What we can do and what we cannot do with fMRI. Nature 453:869–
640 878 Available at: https://www.nature.com/articles/nature06976 [Accessed September
641 25, 2020].

642 Moll J, Krueger F, Zahn R, Pardini M, De Oliveira-Souza R, Grafman J (2006) Human
643 fronto-mesolimbic networks guide decisions about charitable donation. Proc Natl Acad
644 Sci U S A 103:15623–15628 Available at:

www.pnas.orgcgidoi10.1073pnas.0604475103 [Accessed August 3, 2020].

O'Doherty J, Kringelbach ML, Rolls ET, Hornak J, Andrews C (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. Nat Neurosci 4:95–102 Available at: http://neurosci.nature.com [Accessed September 11, 2020].

Op de Beeck HP (2010) Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? Neuroimage 49:1943–1948.

Parkinson C, Sinnott-Armstrong W, Koralus PE, Mendelovici A, McGeer V, Wheatley T (2011) Is morality unified? evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. J Cogn Neurosci 23:3162–3180.

Popal H, Wang Y, Olson IR (2019) A Guide to Representational Similarity Analysis for Social Neuroscience. Soc Cogn Affect Neurosci:1243–1253 Available at: https://academic.oup.com/scan/article-abstract/14/11/1243/5693905 [Accessed July 28, 2020].

Pornpattananangkul N, Zhen S, Yu R (2018) Common and distinct neural correlates of self-serving and prosocial dishonesty. Hum Brain Mapp 39:3086–3103 Available at: https://www.ncbi.nlm.nih.gov/pubmed/29582512.

Shenhav A, Greene JD (2010) Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. Neuron 67:667–677.

Sul S, Tobler PN, Hein G, Leiberg S, Jung D, Fehr E, Kim H (2015) Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. Proc Natl Acad Sci U S A 112:7851–7856.

667 Tricomi E, Rangel A, Camerer CF, Odoherty JP (2010) Neural evidence for inequality-averse
668 social preferences. Nature 463:1089–1091 Available at: www.nature.com/nature.
669 [Accessed August 3, 2020].

670 Tusche A, Tusche XA, Bo A, Kanske XP, Trautwein XF, Singer T (2016) Decoding the
671 Charitable Brain : Empathy , Perspective Taking , and Attention Shifts Differentially
672 Predict Altruistic Giving.

673 Vogt BA (2005) Pain and emotion interactions in subregions of the cingulate gyrus. Nat Rev
674 Neurosci 6:533–544 Available at: www.nature.com/reviews/neuro [Accessed August 6,
675 2020].

676 Weaverdyck ME, Lieberman MD, Parkinson C (2020) Tools of the Trade Multivoxel pattern
677 analysis in fMRI: a practical introduction for social and affective neuroscientists. Soc
678 Cogn Affect Neurosci 15:487–509 Available at:
679 https://academic.oup.com/scan/article/15/4/487/5824852 [Accessed March 30, 2021].

680 Will GJ, Rutledge RB, Moutoussis M, Dolan RJ (2017) Neural and computational processes
681 underlying dynamic changes in self-esteem. Elife 6.

682 Yin L, Hu Y, Dynowski D, Li J, Weber B (2017) The good lies: Altruistic goals modulate
683 processing of deception in the anterior insula. Hum Brain Mapp 38:3675–3690.

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

686 Yoon L, Kim K, Jung D, Kim H (2021) Roles of the MPFC and Insula in Impression
687 Management under Social Observation. Soc Cogn Affect Neurosci Available at:
688 https://academic.oup.com/scan/advance-article/doi/10.1093/scan/nsab008/6101182
689 [Accessed January 29, 2021].

690 Yoon L, Somerville LH, Kim H (2018) Development of MPFC function mediates shifts in
691 self-protective behavior provoked by social feedback. Nat Commun 9:1–10 Available at:
692 http://dx.doi.org/10.1038/s41467-018-05553-2.

693 Zaki J, Cikara M (2015) Addressing Empathic Failures. Curr Dir Psychol Sci 24:471–476
694 Available at: http://journals.sagepub.com/doi/10.1177/0963721415599978 [Accessed
695 October 20, 2020].

696 Zaki J, Mitchell JP (2011) Equitable decision making is associated with neural markers of
697 intrinsic value. Proc Natl Acad Sci U S A 108:19761–19766 Available at:
698 www.pnas.org/cgi/doi/10.1073/pnas.1112324108 [Accessed August 3, 2020].

699

Legends

Figure 1. Dot discrimination task. (A) The overall flow of the experiment. (B) An example of a single trial in the dot-discrimination task. The trial is an example of Both, two-point trial, where dishonesty would earn two points for both the participant and the partner. (C) Violin plot of the mean probability of choosing the wrong answer for each condition.

Figure 2. Univariate analysis results**.** Regions responding differently to different beneficiaries, and amount of points. (A) The RMPFC activation was highest when points were available for Self, and lowest when points were available for Both (The results are displayed at the threshold of $p < 0.05$, FDR corrected at whole-brain level for display purpose). (B) The activity in the RMPFC was higher as more points were available.

Figure 3. Brain fingerprinting approach (A) A pattern classifier was trained to distinguish Self and Other conditions. The trained classifier was then used to forcefully classify Both condition into either of the two classes. (B) Each participant's Self-class Confidence Score (SCCS) was calculated by taking the dot product of the classifier weight map and the participant's Both contrast map.

Figure 4. Correlation with Self-class confidence score (SCCS) for Both condition. A whole-brain regression analysis where the contrast maps of [Point 1 & 2] vs. Point 0 were regressed against the individuals' SCCS shows the clusters with positive correlations in the RMPFC (A, C) and VMPFC (B, D).

724 **Figure 5**. Correlation between the SCCS and the representational similarity between pairs of
725 conditions. In the RMPFC, the SCCS correlated negatively with the degree of pattern
726 similarity between Other and Both conditions (C), but not between Self and Both conditions
727 (A). In the VMPFC, the SCCS correlated positively with the degree of pattern similarity
728 between Self and Both conditions cluster (B), but not between Other and Both conditions (D).
729 Fisher's r-to-z transformed correlation coefficients in the RMPFC (E) and VMPFC (F).
730
731 **Figure 6.** Correlation with Self-class confidence score (SCCS) for Self and Other conditions.
732 (A) Whole-brain regression analyses where the contrast maps of [Point 1 & 2] vs. Point 0
733 were regressed against the individuals' SCCS show the clusters with positive correlations for
734 Self (red) and Other (green) conditions. (B) VMPFC and ventral striatal (VS) activities,
735 showing positive correlations with the SCCS for the Self condition. (C) Anterior insula (AI)
736 and RMPFC activities showed positive correlations with the SCCS for Other condition.
737
738 **Figure 7**. Correlations with RT. (A) The RT difference of dishonest decisions for Other vs
739 Both was positively correlated with the neural similarity in the VMPFC between Self and
740 Both conditions. (B) The RT of dishonest decisions in Both condition correlated with the
741 neural similarity in the VMPFC between Self and Both conditions. (C) The RT of dishonest
742 decisions in Both condition correlated with the ratio of dishonesty in the Both condition.
743
744 **Figure 8.** A schematic diagram of neural signatures in the MPFC associated with increased
745 selfish motivation for Pareto lies. Individuals with higher selfish motivation in Pareto white
746 lies are characterized by increased VMPFC activity when considering dishonesty in all three
747 conditions and increased RMPFC activity when considering dishonesty for Other and Both.
748 In addition, their neural representations in the VMPFC were similar between selfish and

749 Pareto lying opportunities, but those in the RMPFC were dissimilar between altruistic and

750 Pareto lying opportunities.