

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

This Early Release article has been peer-reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

Copyright © 2021 Kim and Kim

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

Selfish Motivation in White Lies

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

2 **Abbreviated Title: Selfish Motivation in White Lies**

3

4 JuYoung Kim¹, Hackjin Kim^{1*}

5 ¹Laboratory of Social and Decision Neuroscience

6 ²Department of Psychology,

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

8

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

10

11 **Number of pages:** 42

12 **Number of figures:** 8

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

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

15 **Acknowledgments:** This work was supported by the National Research Foundation of Korea

16 Grant funded by the Korean Government (NRF-2018S1A3A2075114)

17

18 **ABSTRACT**

19

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

32

33

34 **SIGNIFICANCE STATEMENT**

35

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

43

44 **INTRODUCTION**

45

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

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

63 Several sub-regions of the medial prefrontal cortex (MPFC) serve a crucial role in
64 moral judgment and generation of dishonest behavior. For example, judgments of the
65 dishonesty of a scenario activate the dorsomedial prefrontal cortex (DMPFC) (Parkinson et
66 al., 2011), spontaneous lying engages the subgenual anterior cingulate cortex (Yin et al.,
67 2016), and the ventromedial prefrontal cortex (VMPFC) is involved in deceiving others (Abe
68 et al., 2007) regardless of whether dishonesty is beneficial to the liar or others

Selfish Motivation in White Lies

69 (Pornpattananangkul et al., 2018). Sub-regions of MPFC have been reported to represent self-
70 and other-regarding values differently, where the individual differences in prosociality are
71 expressed as the spatial gradient along the dorsal-to-ventral axis in representing self- and
72 other-regarding values (Sul et al., 2015). Importantly, the rostromedial prefrontal cortex
73 (RMPFC), which includes the pregenual anterior cingulate cortex (Vogt, 2005), is known for
74 computing the values of the outcomes that benefit both self and others (Hutcherson et al.,
75 2015; Sul et al., 2015) and of context-dependent strategic social decisions (Jung et al., 2018;
76 Yoon et al., 2018). It was recently suggested that MPFC subregions are hierarchically
77 organized so that more dorsal regions utilize additional external sensory information from the
78 environment to regulate more ventral subregions that compute intuitive social valuation based
79 on internal bodily signals (Kim, 2020). From this perspective, we predicted that distinctive
80 patterns of activity across MPFC subregions would reflect individual differences in
81 motivation for Pareto lies, given the roles of VMPFC and RMPFC in intuitive social
82 valuation and context-dependent strategic social valuation, respectively.

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

Selfish Motivation in White Lies

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

96 **MATERIALS AND METHODS**

97

98 **Participants**

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

112

113 **Experimental design and statistical analyses**

114 *Experimental Procedure*

115 Participants were given the following overall instruction and a cover story upon arrival. We
116 informed all the participants that the study was about the change in the subjective experience
117 of stressful noise after the depletion of cognitive resources resulting from performing a
118 cognitive task that requires attention. They were to be exposed to an aversive noise for ten
119 minutes and would have to report the stressfulness of the noise after the main task in the MRI
120 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.

146 ***Dot-screen display time calibration procedure***

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

156 ***Dot-discrimination task***

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

Selfish Motivation in White Lies

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

174

175 **Behavioral data analyses**

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

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

189

190 **Neuroimaging procedures and analyses**

191 *FMRI data acquisition and preprocessing*

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

196 240 mm, 80×80 matrix; 36 slices; voxel size = $3 \text{ mm} \times 3 \text{ mm} \times 3 \text{ mm}$). The fMRI blood
197 oxygen-level dependent (BOLD) activity was measured over one functional run, lasting about
198 25 minutes. We acquired the EPI volumes at an oblique angle to the AC-PC line to decrease
199 the impact of susceptibility artifacts in the orbitofrontal cortex. High-resolution T1-weighted
200 (TR = 1900 ms; TE = 2.52 ms; flip angle = 9° ; 256×256 matrix; $1 \times 1 \times 1$ mm in-plane
201 resolution) structural images and diffusion tensor scans (TR = 3000 ms; TE = 70.0 ms; $224 \times$
202 224 matrix; voxel size = $2 \text{ mm} \times 2 \text{ mm} \times 2 \text{ mm}$) were also obtained. The stimuli were
203 presented through an MR-compatible liquid-crystal display monitor mounted on a head coil
204 (refresh rate: 85 Hz; display resolution: 800×600 pixels; viewing angle: 30° horizontal, 23°
205 vertical).

206 We preprocessed the data using the SPM12 (Wellcome Department of Imaging
207 Neuroscience, University College of London, London, UK). Images were temporally
208 corrected for interleaved slice acquisition, and then realigned to the first volume to correct for
209 head motion and a mean image was created for each participant. The realigned images were
210 normalized to the standard Montreal Neurological Institute (MNI) EPI template, resampled to
211 $2 \times 2 \times 2$ mm voxels, and spatially smoothed using a Gaussian kernel with an 8 mm full
212 width at half maximum (FWHM).

213 *1st level univariate analyses*

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

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

228 *2nd level univariate analyses*

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

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

238 A total variation L1 (TV-L1) pattern classifier (Gramfort et al., 2013) was trained to
239 distinguish between neural patterns associated with the opportunities to lie for Self and Other
240 at the moment of decision. The analysis was performed with Nilearn and nlttools library in
241 Python 3 (Abraham et al., 2014). For each beneficiary, representations in the MPFC of the
242 dishonest opportunities were obtained from individual contrasts combining Point 1 and 2
243 conditions and subtracting point 0 condition at decision onset (e.g., [Point 1+ Point 2] – Point
244 0 for Self condition). As the primary aim of this analysis was to identify individuals'
245 motivation when dishonest opportunities were given to gain for Both, we contrasted the

246 conditions in which participants were motivated to lie (i.e., points would be given when
247 lying) with the condition in which participants had no reason to lie (i.e., no point would be
248 given when lying) for each beneficiary. Supporting our rationale for this analysis, the
249 behavioral data showed that participants were induced to lie by the existence, rather than the
250 amount, of available point to be earned. The classifier was first trained on the MPFC activity
251 pattern for Self and Other beneficiary conditions to distinguish between neural patterns
252 associated with the opportunities to lie for Self and Other. Conducting classification using
253 moderately smoothed data is thought to be effective (Op de Beeck, 2010; Hendriks et al.,
254 2017) especially when the objective of the classification is to generalize across subjects
255 (Chang et al., 2015; Weaverdyck et al., 2020). The MPFC binary mask was taken from a
256 meta-analysis segregating MPFC into sub-regions based on the each region's functional co-
257 activation maps (De La Vega et al., 2016). Of the nine MPFC sub-regions, we excluded the
258 supplementary motor area (SMA) and pre-SMA from the mask as activity related to
259 movements or movement control was not considered in this study. Eight-fold nested cross
260 validation was applied where 10 contrast images out of 72 were held out as test data, and the
261 remaining 62 images were used as the training data at each fold. The best performing weights
262 were selected at each fold, and the final classifier weight map was constructed by taking the
263 average of the weights of the overall folds. The MPFC activity pattern of the individual first-
264 level contrast maps of Both condition were entered into the trained classifier to predict the
265 class of each individual's MPFC activity pattern of Both condition (Fig. 3A).

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

Selfish Motivation in White Lies

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

273 *Second-level regression and correlation analyses with the SCCS*

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

277 *Representational similarity analyses*

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

287

288 **RESULTS**

289

290 **Behavioral results**

291 We first tested whether participants were more likely to report incorrectly when points were
292 available, as this suggests dishonesty, and whether such dishonesty is modulated by the
293 beneficiary of the points. We carried out a two-way repeated-measures ANOVA (rmANOVA)
294 to assess the effect of points and beneficiary on the participants' decisions to be dishonest. A
295 significant main effect of point ($F(2,70) = 26.971, p < .001$) was revealed with a significant
296 linear trend ($F(1,35) = 28.380, p < .001$; Fig. 1C) as expected. This suggests that participants
297 were more dishonest as more points were available. The main effect of the beneficiary was
298 also significant ($F(2,70) = 5.078, p = .009$), and behavioral patterns indicated that the
299 participants were generally more dishonest when points were available for Self or Both than
300 for Other. Beneficiary by point interaction was significant ($F(4, 140) = 3.075, p = .018$),
301 implying that each point had a different impact on dishonest decisions depending on the
302 beneficiary. For each pair of the beneficiaries, we ran a two-by-three rmANOVA with
303 beneficiary and point as factors to investigate the cause of the interaction. The analyses
304 revealed that the beneficiary and point interaction was significant for Self and Other ($F(2, 70)$
305 $= 4.894, p = 0.010$), and Other and Both conditions ($F(2, 70) = 3.722, p = 0.029$), but not for
306 Self and Both conditions ($F(2, 70) = 0.278, p = 0.758$). We tested for the difference of 2-way
307 interaction terms between pairs of conditions. The interaction of beneficiary and point was
308 calculated as $(P1 + P2) / 2 - P0$ for each beneficiary, and the difference of the interaction
309 term between each pair of beneficiaries were entered in paired t-tests. The analyses revealed
310 that 2-way interaction of points and beneficiary of Self and Other condition pair was
311 significantly different from the 2-way interaction of Self and Both condition pair ($t(36) = -$

312 2.514, $p = .017$), and the 2-way interaction of Other and Both condition pair was significantly
313 different from Self and Both condition pair ($t(36) = 2.854$, $p = .007$). This indicates a
314 selectively lower dishonesty rate for Other as opposed to Self and Both conditions. The main
315 effect of the point and beneficiary, and the interaction of the two were not significant for the
316 response time (RT) data, but we observed a significant negative correlation between the ratio
317 of dishonest decisions in Both condition and the RT of dishonest choices in Both condition
318 (Fig. 7C, Spearman's $\rho(28) = -.561$, $p = .002$, two-sided). This suggests the individuals who
319 were faster in dishonest decisions for Both were more prone to be dishonest in Both condition.

320

321 **Neuroimaging results**

322 *Univariate analysis result*

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

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

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

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

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

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

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

364 ***Neural evidence for selfish motivation in Pareto lies***

365 Next, we identified neural regions related to the degree of selfish motivation in Pareto lies,
366 which was defined as the self-class confidence scores (SCCS). We calculated the SCCS by
367 taking the signed distance of individuals' Both contrast to the hyperplane separating Self and
368 Other. The SCCS ranged from -2.06 to 2.11 with the mean value of 0.03 and SD of 0.90. The
369 absolute value of the score indicates the certainty of the sample being classified into Self
370 (positive sign) or into Other (negative sign). Thus, the SCCS is assumed to indicate
371 individual differences in the degree of selfish motivation when encountering opportunities to
372 gain from dishonesty for both Self and Other. Individuals' SCCS were then regressed on the
373 contrast map of the Both condition to identify the neural regions uniquely associated with the
374 opportunity for Pareto lies as a function of the degree of selfish motivation. This analysis
375 revealed that the activities in RMPFC ($x = 8, y = 34, z = 14$; Fig. 4A, C), VMPFC ($x = -2, y =$
376 $46, z = -8$; Fig. 4B, D), and precuneus ($x = 10, y = -60, z = 34$) positively correlated with the
377 SCCS.

378 Unlike standard univariate tests, MVPA is now known to be insensitive to inter-
379 subject variability in mean activation across voxels within a region of interest (Davis et al.,
380 2014). Accordingly, we used a representational similarity analysis (RSA) to examine whether
381 multi-voxel patterns in each of the three neural regions associated with the SCCS uniquely
382 encode neural evidence for selfish motivation in Pareto lies. As expected, our analysis
383 revealed that the SCCS correlates *positively* with the degree of similarity of the VMPFC
384 activity pattern between Self and Both conditions (Pearson's $r(36) = .364, p = .029$, two-
385 sided; Fig. 5B), but not between Other and Both conditions (Pearson's $r(36) = .123, p = .475$,
386 two-sided; Fig. 5D). In addition, the SCCS correlates *negatively* with the degree of similarity

387 of the RMPFC activity pattern between Other and Both conditions (Pearson's $r(36) = -.401$, p
 388 $= .013$, two-sided; Fig. 5C), but not between Self and Both conditions (Pearson's $r(36) = -$
 389 $.071$, $p = .681$, two-sided; Fig. 5A). Tests for differences in dependent correlations showed
 390 that the correlation coefficients of Self-Both similarity and Other-Both similarity in RMPFC
 391 cluster are significantly different ($z = 1.651$, $p = .049$, one-tailed; Fig. 5E), and the correlation
 392 coefficients of Self-Both similarity and Other-Both similarity in VMPFC cluster are
 393 marginally different ($z = 1.567$, $p = .057$, one-tailed; Fig. 5F). In the precuneus cluster, the
 394 SCCS showed no correlation with the degree of pattern similarity between Self and Both
 395 (Pearson's $r(36) = .293$, $p = .082$) nor between Other and Both (Pearson's $r(36) = .306$, p
 396 $= .069$).

397 ***Behavioral evidence for selfish motivation in Pareto lies***

398 We examined whether this neural evidence for selfish motivation in Pareto lies can be
 399 validated by behavioral evidence for Pareto lies. Specifically, we examined the difference
 400 between altruistic and Pareto lies as measured by the difference in the proportion of
 401 dishonesty and RT between Other and Both conditions. First, as for the VMPFC cluster, the
 402 degree of similarity between Self and Both conditions in the activity pattern does not
 403 correlate either with the proportion of dishonest choices in Both condition (Pearson's $r(36) =$
 404 $-.066$, $p = .699$) or with the difference in the proportion of dishonest choices between Other
 405 vs. Both conditions (Pearson's $r(36) = .106$, $p = .373$, two-sided). However, the same indices
 406 show a significant *negative* correlation with the RT of being dishonest in Both condition
 407 (Spearman's $\rho(28) = -.492$, $p = .006$, Fig. 7B) and also a significant *positive* correlation
 408 with RT differences between Other vs. Both conditions when being dishonest (Spearman's
 409 $\rho(28) = .463$, $p = .013$, two-sided; Fig. 7A). These findings suggest that those with a high
 410 degree of selfish motivation in lying for Both engage qualitatively different processes
 411 subserving altruistic and Pareto lies, which appears mainly due to their faster intuitive

Selfish Motivation in White Lies

412 responses in Pareto lies. As for the RMPFC and the precuneus clusters, no significant
413 correlation was found between the representational similarity indices and either Other-Both
414 differences in dishonest decisions (RMPFC: Pearson's $r(36) = .136, p = .426$; precuneus
415 Pearson's $r(36) = -.009, p = .957$) or those in RT (RMPFC: Spearman's $\rho(28) = -.016, p$
416 $= .934$; precuneus: Spearman's $\rho(28) = .064, p = .754$).

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

419 We also examined whether and how selfish motivation in Pareto lies is differentially
420 associated with the neural representations in self- and other-benefiting dishonest
421 opportunities. To achieve this, we regressed the SCCS on the contrast map of the Self and
422 Other conditions separating them into two multiple regression analyses. During Self
423 condition, the activities in the VMPFC ($x = 8, y = 44, z = -10$) and the ventral striatum (VS: x
424 $= -16, y = 8, z = -8$) showed significant positive correlations with the SCCS (Fig. 6B). This
425 suggests that as individuals consider opportunities for Both to be closer to opportunities for
426 Self, self-benefiting dishonest opportunities engaged VMPFC and ventral striatum to a larger
427 extent. During the Other condition, a significant positive correlation was observed between
428 individual SCCS and the activities in VMPFC ($x = -6, y = 48, z = -6$), and VS ($x = -18, y = 8,$
429 $z = -2$), similar to the observation made for the Self condition. However, the RMPFC ($x = 6, y$
430 $= 52, z = 16$) and left anterior insula (AI: $x = -26, y = 20, z = -12$) additionally showed
431 significant positive correlations with the SCCS during Other condition (Fig. 6C). These
432 findings indicate that other-benefiting dishonesty additionally engages RMPFC and AI
433 among those Pareto lies that appeared to be primarily driven by selfish motivation.

434 **DISCUSSION**

435

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

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

Selfish Motivation in White Lies

459 gain. Moreover, the degree of pattern similarity in the VMPFC between Self and Both
460 condition predicted faster response times for Pareto lies, the possible indicator of impulsive
461 motivation for earning points by lying. Given the well-known functions of VMPFC in
462 processing reward-predicting information (Knutson et al., 2000; O’Doherty et al., 2001; Kim
463 et al., 2011) and intuitive valuation for decision-making (Shenhav and Greene, 2010; Tricomi
464 et al., 2010; Buckholtz and Marois, 2012; Crockett, 2013; Janowski et al., 2013; Sul et al.,
465 2015; Zaki and Cikara, 2015; Jung et al., 2018), these findings suggest that the increased
466 mean activity in the VMPFC, as well as its specific multivoxel representational pattern that is
467 shared between Self and Both conditions, are the core neural evidence and signatures of
468 selfish motivation in Pareto white lies.

469 Unlike the VMPFC where the mean activity was positively correlated with the SCCS
470 in all three conditions, the higher activity in the RMPFC and AI was positively associated
471 with the SCCS when encountering opportunities for altruistic lies and Pareto white lies, but
472 not for selfish lies. It was recently suggested that the MPFC can be hierarchically organized
473 such that the RMPFC utilizes additional external sensory information from the environment
474 to predict and prevent conflicts occurring in VMPFC tuned to internal bodily signals (Kim,
475 2020). Consistent with this idea, whereas VMPFC is involved in the internalized/intuitive
476 social valuation, RMPFC contributes to the arbitration between internal and external
477 valuation, playing a key role in context-dependent strategic valuation for social decision-
478 making (Tusche et al., 2016; Jung et al., 2018; Yoon et al., 2018; Cutler and Campbell-
479 Meiklejohn, 2019; Fukuda et al., 2019) including sophisticated and socially appropriate
480 expression of self-protective behavior (Kumaran et al., 2016; Will et al., 2017; Yoon et al.,
481 2018) and socially desirable behavior under social observation (Izuma et al., 2010; Jung et al.,
482 2018; Yoon et al., 2021). Based on these theoretical and empirical studies, we can infer that
483 those with higher selfish motivation in Pareto white lies can be characterized by increased

Selfish Motivation in White Lies

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

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

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

Selfish Motivation in White Lies

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

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

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

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

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

551

552

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.

Selfish Motivation in White Lies

- 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->

Selfish Motivation in White Lies

599 18.2019 [Accessed August 3, 2020].

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

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

605 Hendriks MHA, Daniels N, Pegado F, de Beeck HPO (2017) The effect of spatial smoothing
606 on representational similarity in a simple motor paradigm. *Front Neurol* 8:29 Available
607 at: [/pmc/articles/PMC5446978/](https://pubmed.ncbi.nlm.nih.gov/312446978/) [Accessed March 29, 2021].

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

611 Izuma K, Saito DN, Sadato N (2010) The roles of the medial prefrontal cortex and striatum in
612 reputation processing. *Soc Neurosci* 5:133–147 Available at:
613 <http://www.tandfonline.com/doi/abs/10.1080/17470910903202559> [Accessed August
614 10, 2020].

615 Janowski V, Camerer C, Rangel A (2013) Empathic choice involves vmPFC value signals
616 that are modulated by social processing implemented in IPL. *Soc Cogn Affect Neurosci*
617 8:201–208 Available at: <https://academic.oup.com/scan/article/8/2/201/1624534>
618 [Accessed August 3, 2020].

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

Selfish Motivation in White Lies

- 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_](https://www.frontiersin.org/articles/10.3389/fnins.2020.00281?utm_source=researcher_app&utm_medium=referral&utm_campaign=RESR_MRKT_Researcher_inbound)
625 [app&utm_medium=referral&utm_campaign=RESR_MRKT_Researcher_inbound](https://www.frontiersin.org/articles/10.3389/fnins.2020.00281?utm_source=researcher_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:

Selfish Motivation in White Lies

- 645 www.pnas.org/cgi/doi/10.1073/pnas.0604475103 [Accessed August 3, 2020].
- 646 O’Doherty J, Kringelbach ML, Rolls ET, Hornak J, Andrews C (2001) Abstract reward and
647 punishment representations in the human orbitofrontal cortex. *Nat Neurosci* 4:95–102
648 Available at: <http://neurosci.nature.com> [Accessed September 11, 2020].
- 649 Op de Beeck HP (2010) Against hyperacuity in brain reading: Spatial smoothing does not
650 hurt multivariate fMRI analyses? *Neuroimage* 49:1943–1948.
- 651 Parkinson C, Sinnott-Armstrong W, Koralus PE, Mendelovici A, McGeer V, Wheatley T
652 (2011) Is morality unified? evidence that distinct neural systems underlie moral
653 judgments of harm, dishonesty, and disgust. *J Cogn Neurosci* 23:3162–3180.
- 654 Popal H, Wang Y, Olson IR (2019) A Guide to Representational Similarity Analysis for
655 Social Neuroscience. *Soc Cogn Affect Neurosci*:1243–1253 Available at:
656 <https://academic.oup.com/scan/article-abstract/14/11/1243/5693905> [Accessed July 28,
657 2020].
- 658 Pornpattananankul N, Zhen S, Yu R (2018) Common and distinct neural correlates of self-
659 serving and prosocial dishonesty. *Hum Brain Mapp* 39:3086–3103 Available at:
660 <https://www.ncbi.nlm.nih.gov/pubmed/29582512>.
- 661 Shenhav A, Greene JD (2010) Moral judgments recruit domain-general valuation
662 mechanisms to integrate representations of probability and magnitude. *Neuron* 67:667–
663 677.
- 664 Sul S, Tobler PN, Hein G, Leiberg S, Jung D, Fehr E, Kim H (2015) Spatial gradient in value
665 representation along the medial prefrontal cortex reflects individual differences in
666 prosociality. *Proc Natl Acad Sci U S A* 112:7851–7856.

Selfish Motivation in White Lies

- 667 Tricomi E, Rangel A, Camerer CF, O’Doherty 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].

Selfish Motivation in White Lies

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

700 **Legends**

701

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

706

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

712

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

718

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

723

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.

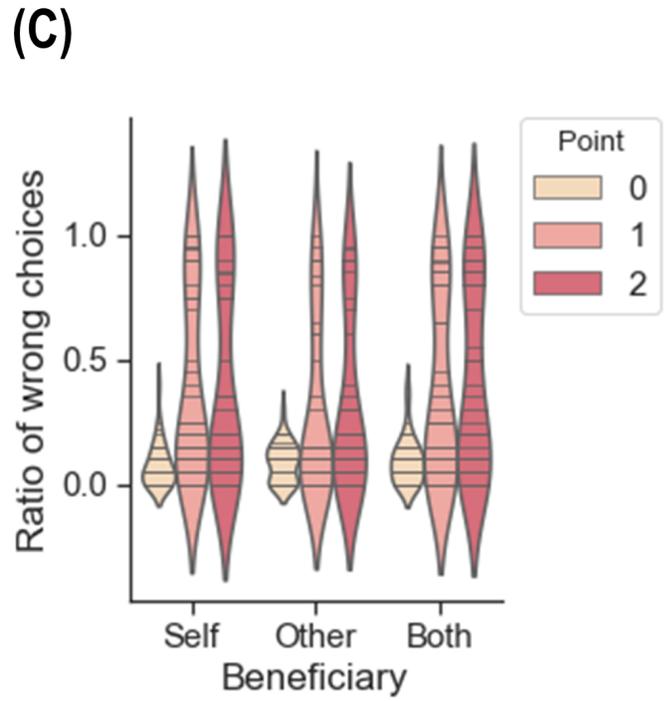
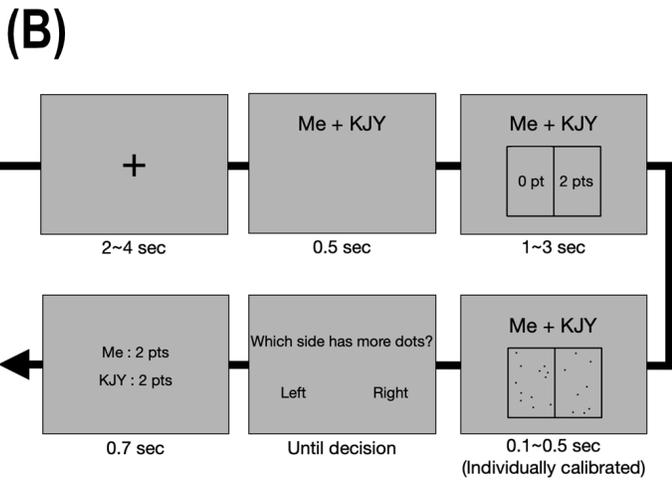
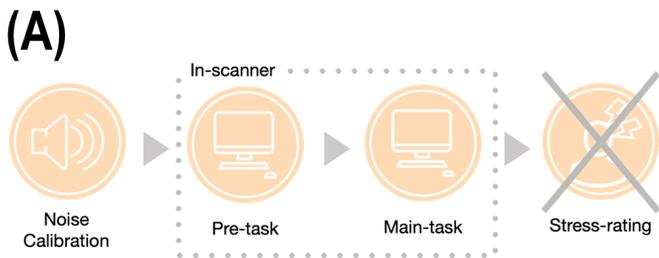
743

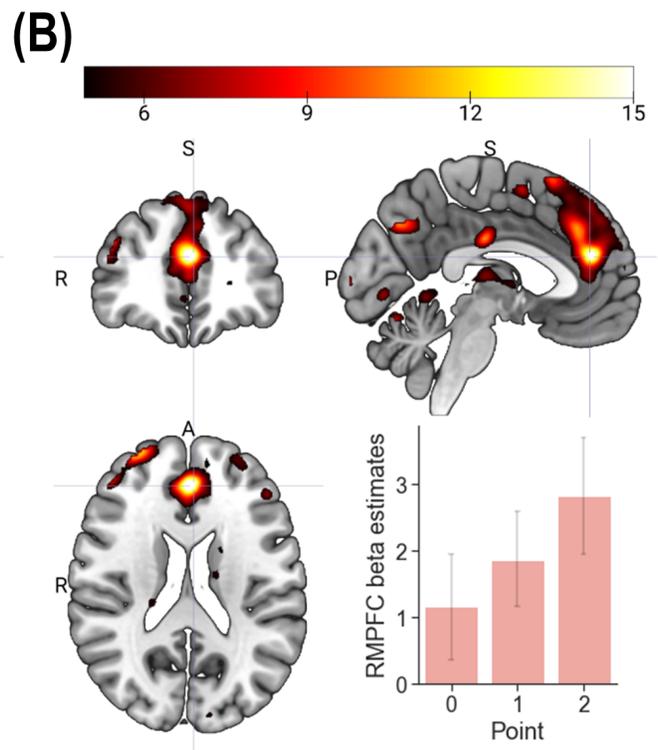
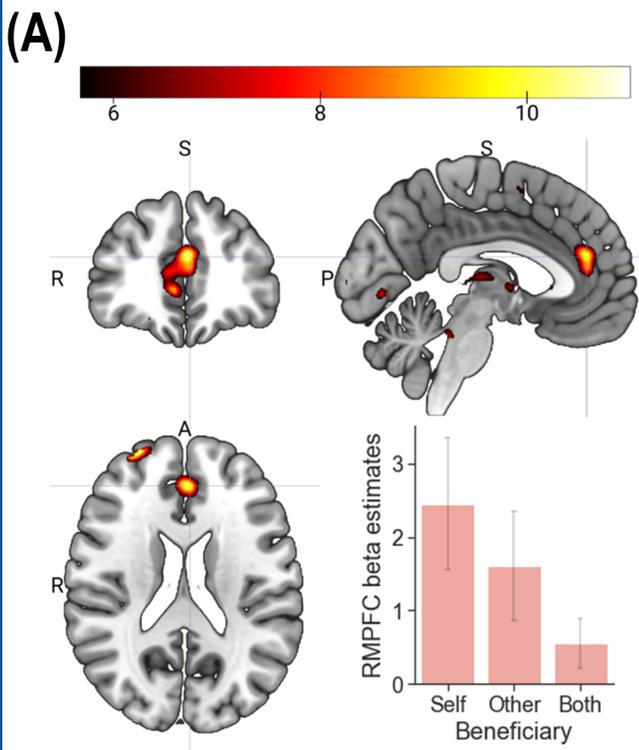
744 **Figure 8.** A schematic diagram of neural signatures in the MPFC associated with increased
745 selfish motivation for Pareto lies. Individuals with higher selfish motivation in Pareto white
746 lies are characterized by increased VMPFC activity when considering dishonesty in all three
747 conditions 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

Selfish Motivation in White Lies

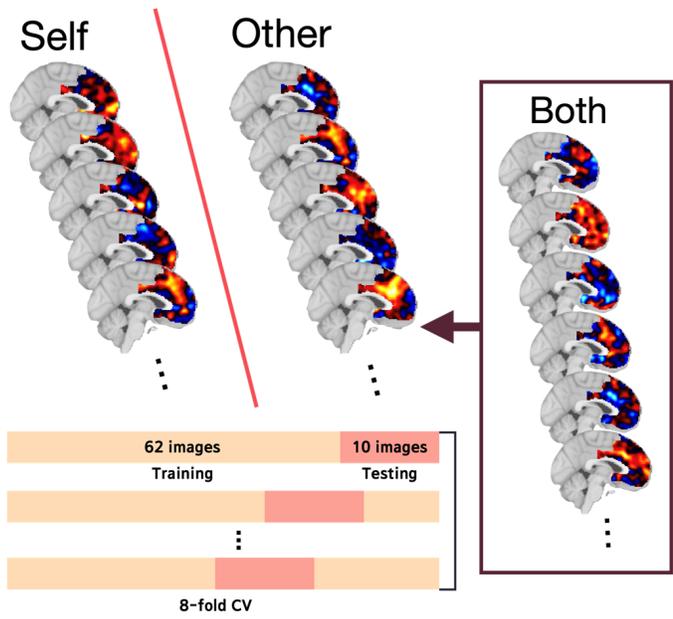
749 Pareto lying opportunities, but those in the RMPFC were dissimilar between altruistic and

750 Pareto lying opportunities.

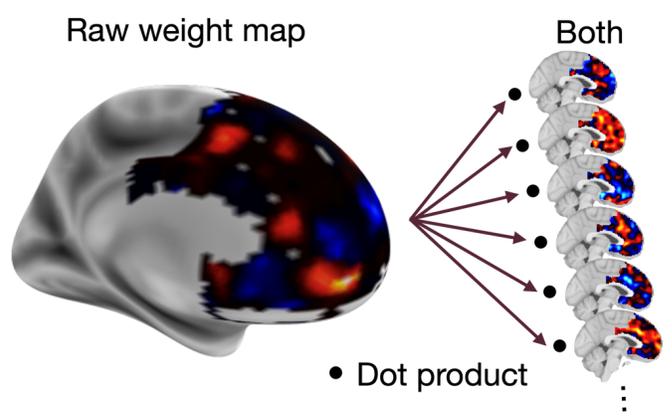




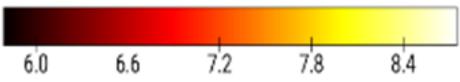
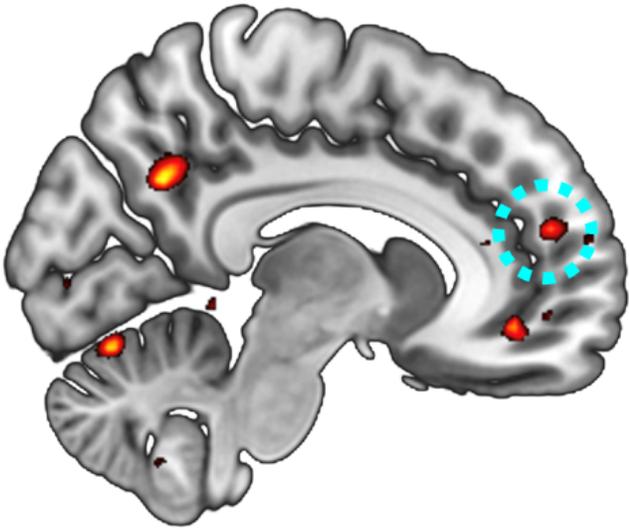
(A)



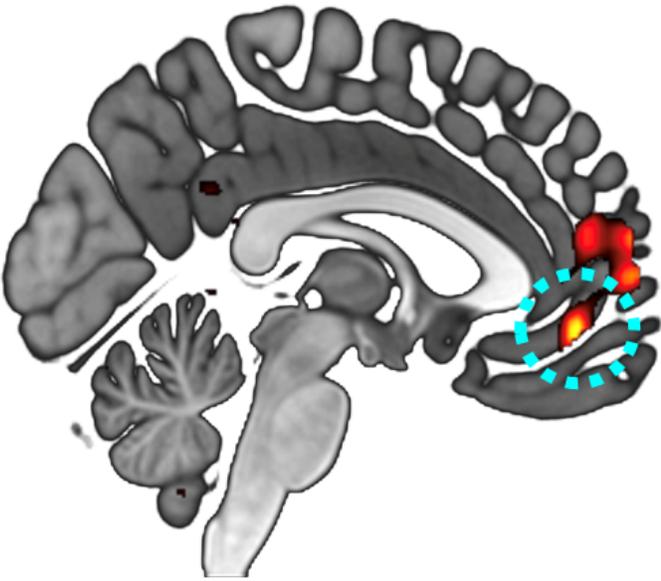
(B)



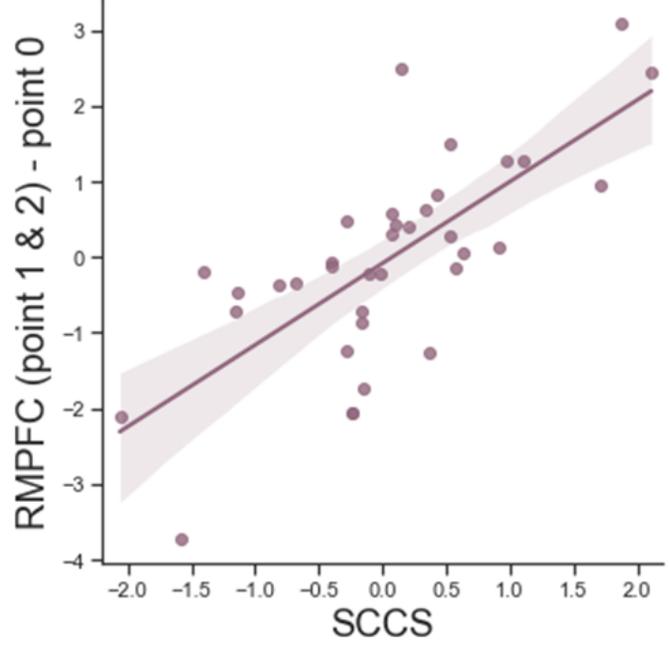
(A)



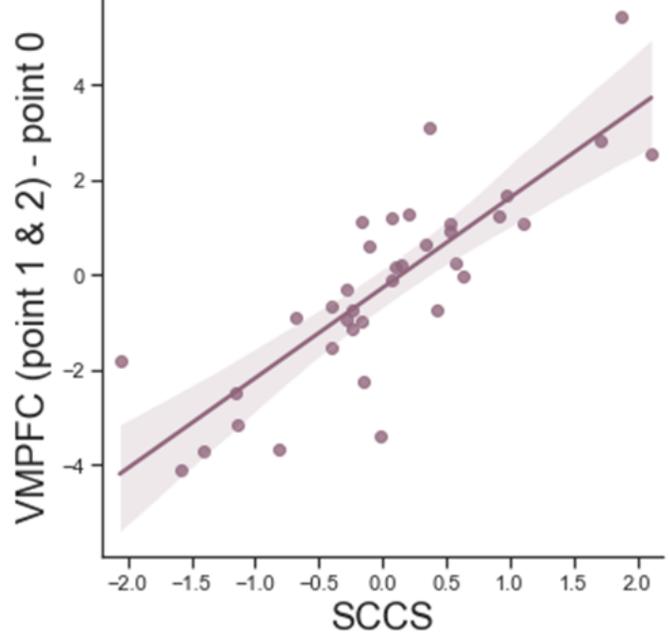
(B)



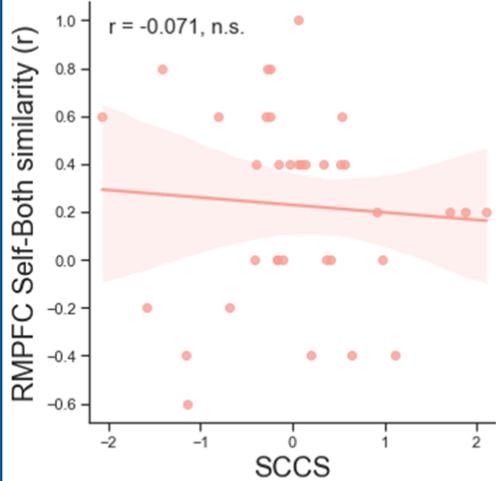
(C)



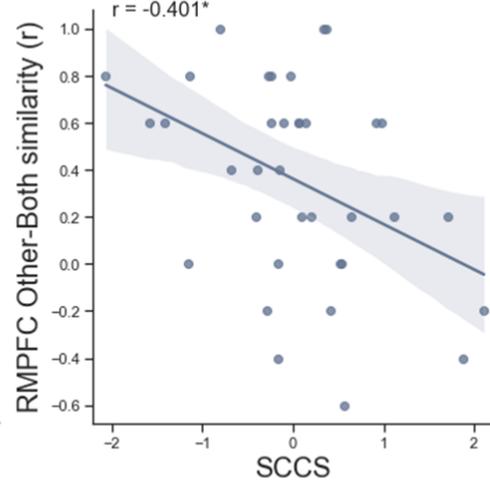
(D)



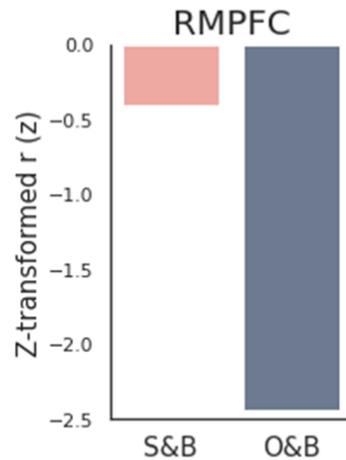
(A)



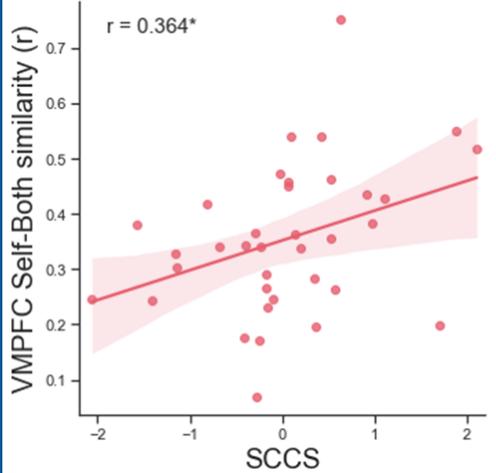
(C)



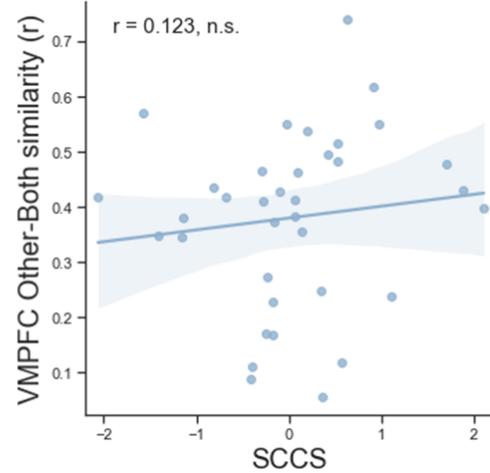
(E)



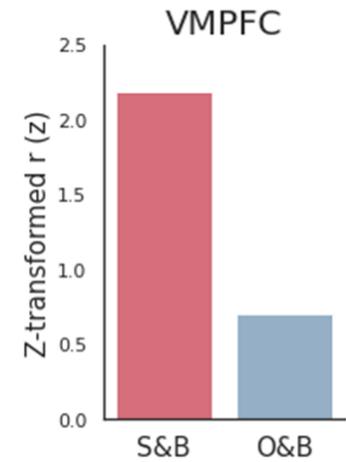
(B)

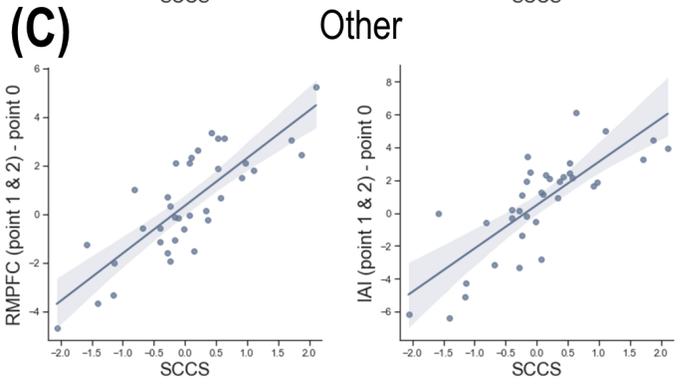
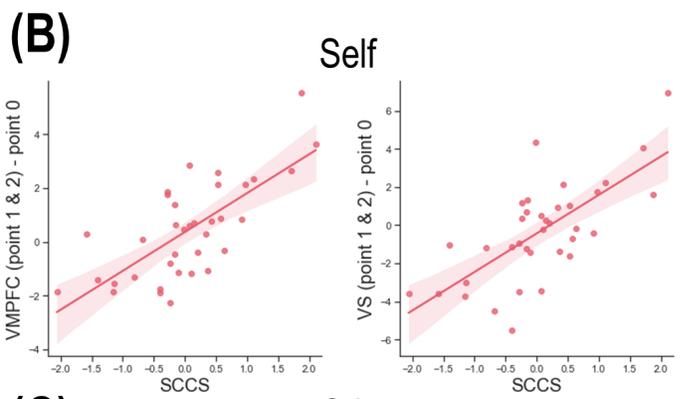
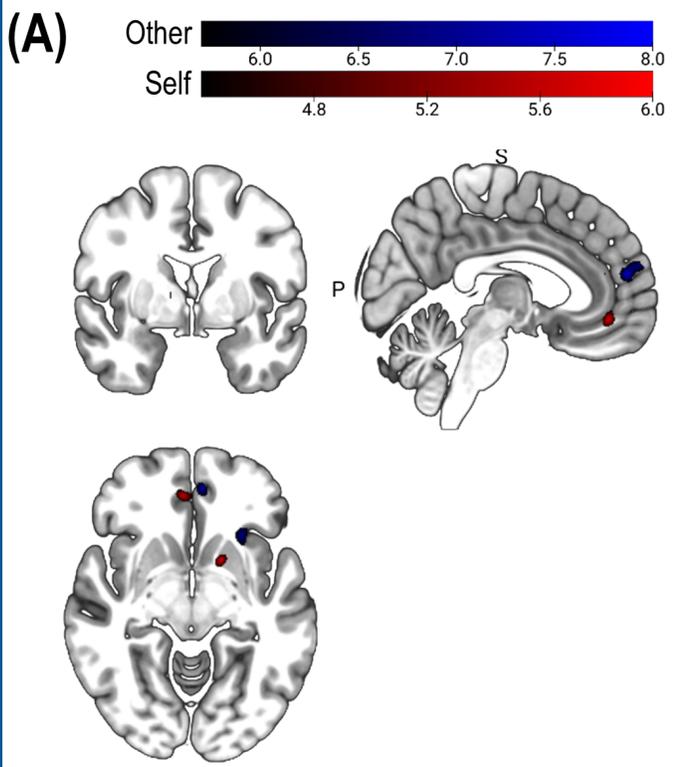


(D)

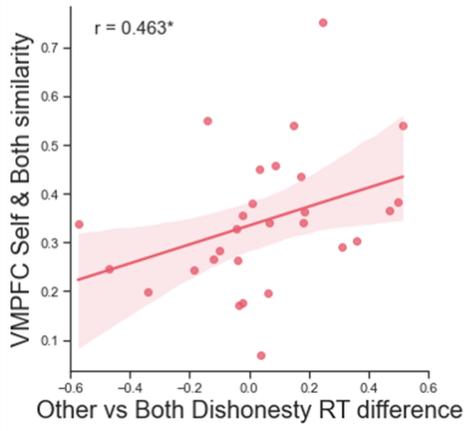


(F)

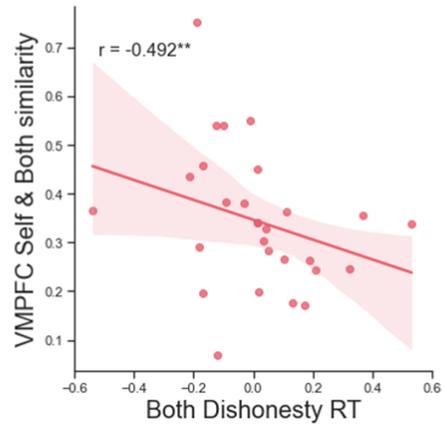




(A)



(B)



(C)

