PNAS
www.pnas.org

Supplementary Information for

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

Sebastian P.H. Speer*, Ale Smidts, Maarten A.S. Boksem
Rotterdam School of Management, Erasmus University, 3062 PA Rotterdam, The Netherlands
Correspondence to: speer@rsm.nl

This PDF file includes:

Supplementary Information
SI References

www.pnas.org/cgi/doi/10.1073/pnas.2003480117

Supplementary Information

Appendix 1 – Visual search task

To further increase the credibility of our cover story on brain processes underlying visual search, we also included the visual search task introduced by Treisman and Gelade (1) at the beginning of our experiment. Specifically, participants were told that the experiment would start with a simple visual search task and then proceed to visual searches in more complex visual stimuli in the second task. In this first task, the goal was to determine whether a specific target was present or absent. In each trial participants were presented with colored letters presented in random locations on the screen. If the target was present, then participants had to press the left button as quickly as possible. If no target was present, then they had to press the right mouse button as quickly as possible. For this task, participants had to search for a green T. Participants were instructed to answer as quickly as possible while still being as accurate as possible. The task took approximately 5 minutes and was also completed in the scanner while localizer scans were obtained to ensure that scanning noise was audible, so participants would believe this task was indeed part of the study**.** This task was not analysed as it was included solely for the purpose of increasing the credibility of our cover story.

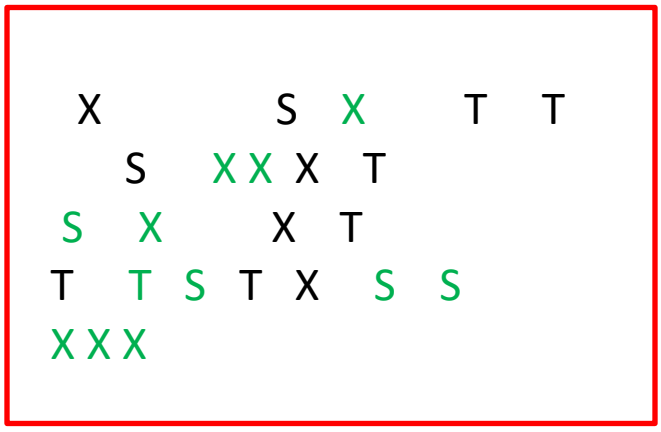

Figure S1. One trial of the simple visual search task. Participants had to indicate whether a green T was among the letters on the screen.

Appendix 2 – Validation of the picture set

Stimuli for the task consisted of 144 Spot-The-Difference image pairs that were downloaded from the Internet. Cartoon images of landscapes containing several objects were selected, to make them engaging and challenging enough for the participants. Landscapes were chosen as they generally satisfied the necessary criteria of containing several different objects, which made the task of spotting differences more challenging and engaging. The stimuli consist of pairs of images that are identical apart from a certain number (1-3) of differences that were created by the experimenter using Adobe Photoshop. Differences consisted of objects added to or removed from the landscape picture or changed colors of objects.

To make sure that participants would be able to find the differences between the images in a reasonable amount of time and to reduce the chance of participants believing that they have seen a difference when they have not (false positives), we ran a pilot study on Amazon's Mechanical Turk with 205 subjects using 180 pictures to test the difficulty to spot the differences between the images and to determine the optimal duration of picture presentation. Participants were presented with cartoon image pairs, presented horizontally next to each other, containing three differences and were asked to click on the differences identified in the image on the right hand side. They were given 15 seconds to give their response. Using the heatmap function provided by Qualtrics, regions of interest were defined around the locations of the differences in the image on the right hand side and response times for each of the clicks were recorded. This allowed us to test whether participants were able to find all differences in an image pair, which differences were particularly difficult to find, and how long it took to identify all differences. Based on the responses of these 205 participants, 36 image pairs that took too long or had differences that were too difficult or too easy to spot, were removed, resulting in 144 images that took 92% participants less than 6s to find all three differences (M=5.4s, SD =1.5s). This high success rate of finding the three differences also points to the low ambiguity of the differences, which reduces the chance of false positives. While we cannot completely rule out the chance of false positives it has to be noted that these false positives are unlikely to contribute anything other than noise to the data.

Appendix 3 – Regions extracted for ROI analyses

Depicted here are the tables showing the regions extracted from Neurosynth.

Table S1. ToM and Cognitive Control masks link for download

Network	Studies	Date of	Link to download
Self Referential	166	03.06.2019	http://neurosynth.org/analyses/terms/self%20referential/
Cognitive Control	598	03.06.2019	http://neurosynth.org/analyses/terms/cognitive%20control/
Reward Anticipation	92	03.06.2019	https://neurosynth.org/analyses/terms/reward%20anticipation/

Appendix 4 – Cluster statistics for the second-level results for cheatable vs non-cheatable trials

Table S2. Regions more activated during cheatable trials as compared to non-cheatable trials for honest participants as compared to cheaters

Region	cluster_id	peak_x	peak_y	peak_z	peak_value	volume_mm
PCC	1	-9	-57	23.79	492.67	10014
R TPJ	2	45	-60	23.79	445.75	4138
Hippocampus	3	24	-18	-18.33	440.28	3632
(v)MPFC	4	-6	54	-4.29	388.30	3538
Cerebellum	5	0	-54	-60.45	379.48	3222
MFG	6	-30	24	44.85	407.82	3032
Cerebellum	7	-27	-48	-25.35	429.12	2811
Left Frontal Pole	8	-18	39	44.85	421.57	2337
MPFC	9	-6	30	6.24	39.30	2053
L TPJ	10	-45	-69	23.79	38.96	1674
R Postcentral Gyrus	11	30	-42	65.91	382.64	1547
R Supramarginal Gyrus	12	66	-30	27.3	448.58	1263
L Supp motor area	13	0	0	48.36	365.03	1232
L C	14	-18	-42	-49.92	411.61	1105
R Cerebellum	15	12	-45	-11.31	429.17	1105
R Hippocampus	16	21	-39	6.24	466.65	1105
L OFC	17	-42	36	-14.82	431.2	1042

Appendix 5 – Cluster statistics for the second-level results for cheated vs honest decisions

Table S3. Regions more activated during honest decisions as compared to cheated decisions for cheaters than for honest participants

Region	cluster_id	peak_x	peak_y	peak_z	peak_value	volume_mm
L IFG	1	-46	21	-5	41	4156
R ACC	2	7	36	21	384	2797
L ACC	3	-7	41	7	450	1922
R Insula	4	37	26	-6	387	762
L Frontal Pole	5	-34	62	5	476	704
L Supp Motor Area	6	-11	23	63	398	639
L Nacc	7	-14	19	-7	372	326
L SFG	8	-4	20	43	356	272
R Cingulate Gyrus	9	1	-28	29	330	237
R Angular Gyrus	10	54	-51	45	331	200

Here we also find the left Nacc to be activated, which seems inconsistent with the other findings. However, it has to be noted that these activations, including the Nacc, were further tested in the trial-by-trial analysis. In this trial-by-trial analysis we investigate which of the previously identified regions is most important in predicting trial-by-trial cheating. This analysis includes the NAcc as well as the ACC and the IFG, and there the Nacc was not found to be a significant predictor of the decisions to be honest for cheaters. This suggests that the cluster reported here may have been a false positive. An alternative explanation could be that cheaters experience a warm glow effect (2), which proposes that people behave selflessly or morally because they are compensated by the warm glow of knowing they have acted prosocially. Honest participants may intuitively act honesty without further thinking about it whereas cheaters may do so more rarely and when they do so they experience the warm glow which is represented in the Nacc. This is, however, very speculative and further research would be needed to confirm these speculations.

Appendix 6 – Cluster statistics for the second-level results of the parametric modulation analysis for the level of reward

Table S4. Regions parametrically modulated by level of reward during the level of difficulty phase of the Spot-The-Difference task

Region	cluster_id	peak_x	peak_y	peak_z	peak_value	volume_mm
Left Cuneus	1	-9	-78	16.77	546	1611
R Nacc	2	12	12	-0.78	493	1232
L Nacc	3	-21	15	-0.78	47	568
L Cuneus	4	-6	-96	27.3	414	315

Appendix 7 – Levels of engagement during visual search

In order to test whether our findings may be confounded by different levels of engagement during the visual search phase, we tested whether there were differences in neural activation during the visual search phase between more honest participants and cheaters. First, we ran a univariate analysis in which we contrasted neural activity during the visual search against baseline activation. The analysis revealed that a large cluster in the visual cortex showed higher activation during search as compared to baseline activation, which is expected as participants were engaged in visual search. In addition, several regions related to working memory, cognitive processing and navigation, such as the dmPFC and the MFG were more strongly activated during visual search (see Table S5 for table with cluster statistics).

To explore whether there are individual differences in level of engagement during visual search, participants' cheat count was added as a group level covariate. The whole brain analysis revealed that there are no significant differences between more honest participants and cheaters during the visual search phase. In addition, we also tested whether differences in neural activation during visual search between cheatable and non-cheatable trials were more strongly expressed in cheaters or honest participants. In order to do so, a univariate analysis was run in which we contrasted neural activation during visual search in cheatable trials against activation during visual search in non-cheatable trials. Again, these contrast maps were then correlated with cheat count on the group level. The whole brain analysis did not reveal any significant effects. These findings suggest that there are no significant differences in level of engagement or motivation during visual search between more honest participants and cheaters.

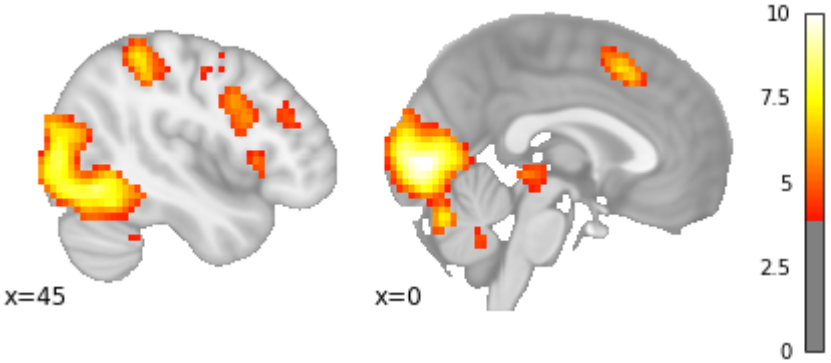

Figure S2. The visual cortex, dMPFC and left and right dlPFC are more activated during visual search as compared to baseline

Table S5. Regions more activated during visual search as compared to during rest

Region	cluster_id	peak_x	peak_y	peak_z	cluster_mean	volume_mm
Occipital Cortex	1	0	-84	2.73	643.202	301369
dmPFC	2	0	15	48.36	514.447	16490
MFG	3	24	6	51.87	515.819	6981.39
R dlPFC	4	45	6	30.81	488.557	6160.05
R Insula	5	30	27	-0.78	553.421	5907.33
R dlPFC	6	-48	0	30.81	448.594	3601.26
L Insula	7	-33	21	-0.78	514.818	3064.23
Cerebellum	8	-18	-42	-46.41	490.497	1769.04
R lPFC	9	51	36	27.3	439.172	1674.27
Cerebellum	10	-30	-69	-53.43	494.043	663.39

To corroborate this neural evidence with behavioral data, we tested whether there were significant differences in accuracy on the simple visual search task (see Appendix 1) between honest participants and cheaters. The analysis revealed that there were no significant differences between honest participants and cheaters in accuracy on the simple visual search task ($t = 1.17$; $p = 0.25$; participants were categorized in groups by median split). Assuming participants were honest on three differences trials, we could also compare the behavioral accuracy between cheaters and honest participants on the Spot-The-Difference task. We performed this analysis and found no significant differences ($t=1.54$, $p=0.16$) in how often cheaters or honest participants (as categorized by median split) reported to have found three differences when there were actually three differences. Collectively, these findings suggest that there were no significant differences in levels of engagement during the visual search of the Spot-The-Difference task.

Appendix 8 – Factor analysis to confirm validity of networks

To test whether the regions we are analyzing indeed belong to three separate networks, we conducted an exploratory factor analysis with promax rotation (3), which is an oblique rotation method which allows for correlation between latent factors. Specifically, the goal of this factor analysis was to determine the most important latent factors underlying all the regions resulting from our conjunction analyses, namely the left IFG and ACC (cognitive control network), the PCC, bilateral TPJs and MPFC (self-referential network), and the bilateral Nacc (reward network).
We used the single trial activations obtained as explained above by fitting a model that includes a separate regressor for each trial from each of the regions as input for the factor analysis. Before conducting the factor analysis, we first checked whether the regions intercorrelated at all using Bartlett's test of sphericity, which tests the observed correlation matrix against the identity matrix. Bartlett's test indicated that the null hypothesis can be rejected and there is significant correlation between variables justifying a factor analysis ($\chi 2 = 10582$, $p<0.001$). In addition, the Kaiser-Meyer-Olkin (KMO) test was conducted which determines the adequacy of the observed variables by estimating the proportion of variance among all the observed variables. The KMO test revealed an overall estimate of 0.69 which indicates that the observed variables are adequate for a factor analysis.
Next, we determined the number of factors with the help of the Kaiser criterion (choosing factors with an eigenvalue >1). This resulted in three latent factors, where the first factor represented the self-referential thinking network with the bilateral TPJs, PCC and the MPFC loading highly on this factor. The second factor clearly represents the reward network as only the bilateral Nacc show high factor loadings. Lastly, the third factor clearly represents the cognitive control network as only the ACC and the left IFG load highly on this component. This exploratory factor analysis clearly indicates that the regions of interest used in our trial-by-trial and functional connectivity analysis indeed belong to three separate networks.

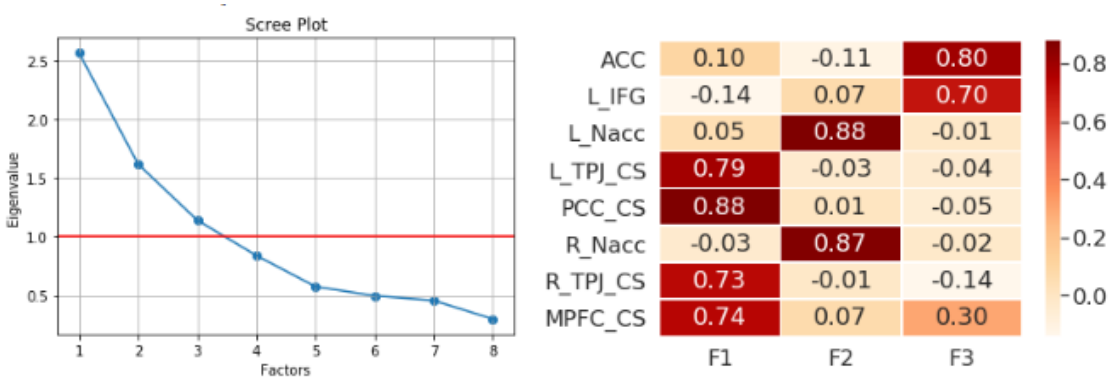

Figure S3. Left: Scree plot showing the the eigenvalues for each factor. Right: the loadings for each of the factors

Appendix 9 – Classifying cheaters versus honest participants and predicting trial by trial cheating

Due to the fact that we found that we could classifiy cheaters and honest participants based on the functional connectivity patterns during decision-making, we wanted to see whether average activation within a subject in the ROIs from the three networks of interest (cogntive control, reward & self referential thinking) could be used to classify participants as cheaters or honest participants (categorized by median split). In order to do this, we average the trial-by-trial estimates within participants, resulting in one observation for each subject, which represents the average activation in each ROI across the whole task.
In order to test this, we employed a support vector classifier (4, 5) with a linear kernel (C=1), trained on average activations in the ROIs of each participant to determine whether a participant was a cheater or an honest participant (categorized by median split). To avoid overfitting and inflated prediction accuracy (6), this was done using 8-fold cross validation. Significance was estimated using permutation testing (N=5000). The classification analysis revealed that we could significantly classify an unseen participant as a cheater or an honest participant based on the average activation in the ROIs (F1=70%, AUC=77%, $p<0.05$). Using activations from honest trials only an even higher classification accuracy was found (AUC=84%, $p<0.05$). Classification was not significant using cheated trials only.

We also tested whether combining the model predicting trial-by-trial cheating, using the trial-by-trial activation from the ACC and IFG, could be improved by adding the output from the model classifying cheaters versus honest participants based on the participants connectivity patterns. In order to increase statistical power, instead of using a support vector machine trained on participants that were median split on cheatcount, we used a support vector regression approach to predict the cheatcount of an unseen participant based on participants' connectivity patterns. This allowed us to use the full range of the participants' cheatcounts. Specifically, as in the model reported in the manuscript, we used 8-fold cross validation to train a support vector regression (SVR) model on the connectivity patterns of our participants to predict the cheatcount of an unseen participant. The predictions from the SVR model correlated significantly with the cheatcount ($r=0.73$, $p<0.05$), demonstrating the predictive accuracy of the SVR model. In a direct model comparison, adding the output from the SVR to the multilevel model with ACC and left IFG led to a significantly improved fit ($\chi^2=14.1$, $p<0.05$). However, when testing the model using 8-fold cross validation, no substantial improvement in predictive accuracy was found (AUC=79%, F1=85% as compared to AUC=76%, F1=89%).This could be due to the intercept of the multilevel model already accounting for individual differences in moral default that are similarly explained by the connectivity in the self-referential thinking network.

To test this conjecture, we also trained support vector machines without an intercept capturing individual differences, on the trial-by-trial data with the activity from the control regions and with or without the output from the connectivity model. This analysis revealed that when using only the control regions, a considerably lower predictive accuracy (AUC=68%) was found as compared to the model with the output from the connectivity model included (AUC=75%). It can thus be concluded that the intercept in the multilevel model indeed captures individual differences in moral default, that are also explained by the output of the model trained on connectivity patterns. In this sense, adding output from the connectivity model increases the interpretability of the model as individual variation in moral default is explicitly captured by variation in connectivity between regions in the default mode network.

Alternatively, or additionally, the ACC, which is already included in the model predicting trial-by-trial cheating, may encode individual differences in moral default that are similarly captured by connectivity within the self-referential thinking network. Whereas we found that higher activity in the IFG increases the probability of cheating in honest participants and decreases the probability of cheating for cheaters, no such effect was found for the ACC. The ACC has been frequently associated with conflict monitoring and conflict detection (7) and may encode individual differences in moral default to some extent. Stated differently, the extent to which honest participants monitor

and detect moral conflict may differ from cheaters and may reflect individual differences in moral default.

Appendix 10 – Example image pairs for the Spot-the-Difference task

To provide a better sense of how difficult it was to spot the difference between images if there were indeed three differences, three example pairs are shown. For the sake of space, image pairs are presented horizontally (next to each other), whereas in the actual Spot-The-Difference task the image pairs were presented vertically (on top of each other). All images are also available in the publicly available repository.

Figure S4. Example image pair used in the Spot-The-Difference task. The right image contains a dragonfly (top left) a red sun with a smiley (top right) and a red flower (middle right), which are not present in the left image.

Figure S5. Example image pair used in the Spot-The-Difference task. The right image contains a hot air balloon (top left) a red flag (centre) and a purple building(right), which are not present in the left image.

Figure S6. Example image pair used in the Spot-The-Difference task. The right image contains a rabbit (left) a monkey (center) and a wooden box (right), which are not present in the left image.

SI References

1. Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*, 97-136.
2. Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, *100*(401), 464-477.
3. Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, *17*(1), 65-70.
4. Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI)"brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, *19*(2), 261-270.
5. Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine learning*, *57*(1-2), 145-175.
6. Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274-290.
7. Carter, C. S., & Van Veen, V. (2007). Anterior cingulate cortex and conflict detection: an update of theory and data. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 367-379.