Unethical amnesia brain: Memory and metacognitive distortion induced by dishonesty

Xinyi Julia Xu (yc27301@connect.um.edu.mo)
Centre for Cognitive and Brain Sciences and Department of Psychology, University of Macau
Macau, China

Dean Mobbs (dmobbs@caltech.edu)
Division of the Humanities and Social Sciences, California Institute of Technology
California, USA

Haiyan Wu* (haiyanwu@um.edu.mo)
Centre for Cognitive and Brain Sciences and Department of Psychology, University of Macau
Macau, China

Abstract

Unethical actions and decisions may distort human memory in two aspects: memory accuracy and metacognition. However, the neural and computational mechanisms underlying the metacognition distortion caused by repeated dishonesty remain largely unknown. Here, we performed two fMRI studies, including one replication study, with an information-sending task in the scanner. The main moral decision task in the scanner involves cumulative history response (CHR) and reward as two main factors, combined with a pre-scan and post-scan memory test together with mouse tracking. With multiple dimensions of metrics to measure metacognition, we test whether the inter-subject metacognition change correlates with how participants trade off consistency and reward. We find that the compression of representational geometry of reward in the orbitofrontal cortex (OFC) is correlated with both immediate and delayed metacognition changes. Also, the functional connectivity between the dorsolateral prefrontal cortex (DLPFC) and the left temporoparietal junction (lTPJ) under dishonest responses can predict both immediate and delayed metacognition changes in memory. These results suggest that decision-making, emotion, and memory-related brain regions together play a key role in metacognition change after immoral action, shedding light on the neural mechanism of the complex interplay between moral decisions, cognitive processes, and memory distortion.

Keywords: metacognition; dishonesty; decision-making

Introduction

Studies on morality and memory have revealed that people forget the instances where they lied or made selfish choices(Carlson, Maréchal, Oud, Fehr, & Crockett, 2020). One leading interpretations stem from cognitive dissonance theory(Aronson, 1969), which suggests that discrepancies between truth and conflicted responses lead to psychological discomfort. Indeed, substantial evidence suggests that forgetting is one way of cognitive dissonance reduction(McGrath, 2017). However, when participants have time to forget about the dissonance, a decline in its level is observed (Elkin & Leippe, 1986). The results raise an interesting question about whether repeated immoral behaviors lead to forgetting, with false memories that can not be discriminated against.

In the present study, with a complementary manipulation of pre-task and post-task measures, and with three conditions in the moral decision(see Figure 1), we investigate (i) whether moral decisions lead to accuracy and metacognition change, and (ii) if so, how this change is associated with cognitive control and memory neural system. We predict the influence of immoral decisions on memory metacognition on memory would be evident by decoupling respective contributions of reward and CHR to behavior and its neural basis.

Methods

Metacognition quantification

Representational similarity analysis with parameterized model Unsmoothed beta maps of different conditions were used. Conditions were defined according to the value of CHR differences (-7 to 7), reward differences (-8, -6, -4, -2, 2, 4, 6, 8) and moral responses (dishonesty as 1 and honesty as 0) were used to construct model RDMs (distances between pairwise conditions according to the value of three factors). Neural RDMs were calculated by the dissimilarity between pairwise conditions using beta maps.

We fit the parameterized model to neural RDMs. We yielded six parameters: $compression_{con}^{lie}$, $compression_{rew}^{lie}$, $compression_{con}^{hon}$, and $compression_{rew}^{hon}$ controlled for the compression along consistency and reward dimension under dishonesty and honesty responses respectively; the $rotation$ parameter controlled for the response-dependent rotation of the variable axes (consistency and reward) from native space into the reference frame of the response; the context $offset$ parameter controlled for the parallel distance between honesty and dishonesty responses.

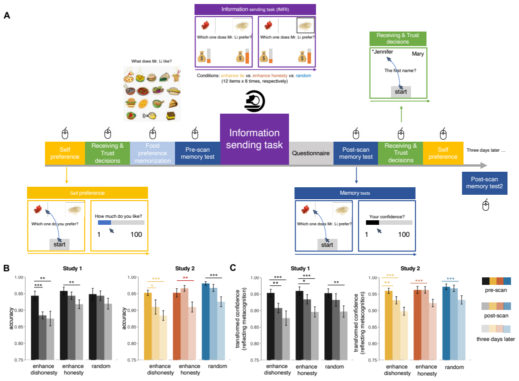

Figure 1: Paradigm and behavioral results. **(A)** There were 9 parts (8 parts for day 1 and 1 part for three days later) for the whole study. Participants performed a food memory task and a food preference task before and after the Information Sending Task (IST). IST was conducted in the fMRI scanner, where participants were asked to pass the preference information of Mr. Li to the next participant (the receiver) with the consideration of reward units in four scan sessions. We manipulated reward unit difference (referred to as "reinforcer" in this paper) between two items and kept them lying about specific items (repeated dishonesty) or making truthful responses with 4 sessions of fMRI scanning. **(B)** The accuracy and **(C)** transformed confidence (using the quadratic scoring rule to quantify metacognition level, see *Methods*) in pre-scan, post-scan and three-day-later memory tasks. blackThe transparency of color represents the period of memory tests.

Results

Compressional neural geometry of consistency and reward under dishonesty and dishonesty responses

Neural geometry of OFC was in Figure c, in which the grid of dishonesty was rotated and reward was compressed. We conducted Spearman's rank-order correlations between inter-subject metacognition change and the compression ratio of dishonesty responses. It showed significant correlations between compression ratio under dishonesty responses and metacognition change of pre-scan to post-scan test in SMA, OFC, lTPJ, PCC and insula (SMA: $rho = 0.10$, $p = 0.045$; OFC: $rho = 0.16$, $p = 0.0025$; lTPJ: $rho = 0.18$, $p < 0.001$; PCC: $rho = 0.16$, $p = 0.0027$; insula: $rho = 0.14$, $p = 0.0073$). For metacognition change from pre-scan to three-days-later test, only OFC was found significantly correlated with the compression ratios (OFC: $rho = 0.27$, $p < 0.001$).

Functional connectivity between DLPFC and lTPJ was associated with metacognition change

Playing a key role in cognitive control and dishonesty(Speer, Smidts, & Boksem, 2022), DLPFC was selected as a hub in functional connectivity(FC) analysis. Results showed significant FCs between DLPFC and PCC, insula and hippocampus (Figure **??**A, left panel), echoing the involvement of DLPFC in memory change. Further, we examined whether there were task-related FCs between DLPFC and other ROIs that could predict the degree of metacognition change. We used inter-subject RDMs and implemented linear regression to explore the extent to which FCs between DLPFC and other ROIs could predict metacognition change variation. The results showed that the FCs between DLPFC and lTPJ significantly predicted immediate and delayed metacognition change (immediate: β = 1.31, $p < 0.001$, 95% CI from 0.69 to 1.92; delayed: β = 0.69, $p = 0.03$, 95% CI from 0.078 to 1.29;).

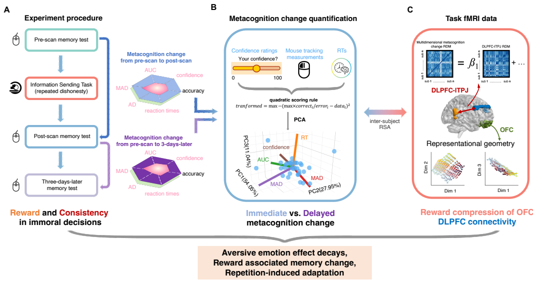
A
Experiment procedure
Pre-scan memory test
Information Sending Task (repeated dishonesty)
Post-scan memory test
Three-days-later memory test
Metacognition change from pre-scan to post-scan
AUC
confidence
MAD
accuracy
AD
reaction times
Metacognition change from pre-scan to 3-days-later
Reward and Consistency in immoral decisions
B
Metacognition change quantification
Confidence ratings
Mouse tracking measurements
RTs
Your confidence?
0
100
quadratic scoring rule
$tranformed = \max - (\max \times correct_i/error_i - data_i)^2$
PCA
RT
confidence
AUC
MAD
PC3(11.04%)
PC1(54.05%)
PC2(27.95%)
Immediate vs. Delayed metacognition change
inter-subject RSA
C
Task fMRI data
Multidimensional metacognition change RDM
DLPFC-lTPJ RDM
$= \beta_1$
+ ...
DLPFC-lTPJ
OFC
Representational geometry
Dim 2
Dim 3
Dim 1
Reward compression of OFC
DLPFC connectivity
Aversive emotion effect decays,
Reward associated memory change,
Repetition-induced adaptation

Figure 2: Methods Summary: Neural mechanism of metacognition change induced by repeated dishonesty.
(A) We recorded RTs, mouse tracking indices, and self-reported confidence in memory tests. **(B)** Metacognition change manifests on multiple dimensions of response in the pre- and post-scan memory tasks. These measurements all reflected metacognition levels after being transformed by the **quadratic scoring rule**(von Holstein, 1970; Fleming, Van Der Putten, & Daw, 2018; Rollwage et al., 2020; Carpenter et al., 2019). They jointly represent the immediate and delayed metacognition change after orthogonalization using PCA. **(C)** We correlated the pairwise distance among participants with measurements from neural analysis (ie. the reward compression gained in neural representational geometry; functional connectivity).

Acknowledgments

This work was mainly supported by the Science and Technology Development Fund (FDCT) of Macau [0127/2020/A3, 0041/2022/A], the Natural Science Foundation of Guangdong Province (2021A1515012509), Shenzhen-Hong Kong-Macao Science and Technology Innovation Project (Category C) (SGDX2020110309280100), MYRG of University of Macau (MYRG2022-00188-ICI) and the SRG of University of Macau (SRG2020-00027-ICI).

References

Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. In *Advances in experimental social psychology volume 4* (p. 1–34). Elsevier. Retrieved from `http://dx.doi.org/10.1016/S0065-2601(08)60075-1` doi: 10.1016/s0065-2601(08)60075-1

Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., & Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature communications*, *11*(1), 1–11.

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*(1), 51.

Elkin, R. A., & Leippe, M. R. (1986). Physiological arousal, dissonance, and attitude change: evidence for a dissonance-arousal link and a" don't remind me" effect. *Journal of personality and social psychology*, *51*(1), 55.

Fleming, S. M., Van Der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature neuroscience*, *21*(4), 617–624.

McGrath, A. (2017). Dealing with dissonance: A review of cognitive dissonance reduction. *Social and Personality Psychology Compass*, *11*(12), e12362.

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature communications*, *11*(1), 2634.

Speer, S. P., Smidts, A., & Boksem, M. A. (2022). Cognitive control and dishonesty. *Trends in Cognitive Sciences*, *26*(9), 796–808.

von Holstein, C.-A. S. S. (1970). Measurement of subjective probability. *Acta Psychologica*, *34*, 146–159.