International Journal of Clinical and Health Psychology 24 (2024) 100511

Contents lists available at ScienceDirect

International Journal of Clinical and Health Psychology

journal homepage: www.elsevier.es/ijchp

Neural impacts of personality on deception for applications of deception detection

Xiaoling Zhang [a,b], Xinyue Zhang [a], Xiumei Chen [a], Qianjin Feng [a,c,d], Meiyan Huang [a,c,d,*]

[a] School of Biomedical Engineering, Southern Medical University, Guangzhou, 510515, China
[b] Zhuhai Institute of Translational Medicine, Zhuhai People's Hospital, Zhuhai Hospital Affiliated with Jinan University, Zhuhai, 519000, China
[c] Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, 510515, China
[d] Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou, China

ARTICLE INFO

Key words:
Functional near-infrared spectroscopy
Deception detection
Personality
Transformer

ABSTRACT

Background: Although universal, deception may be a sign of certain mental disorders and may impede the effectiveness of clinical intervention. However, individual difference exists in deception, and its underlying neural mechanisms remain unclear, thereby limiting the applicability of deception detection for clinical diagnosis.

Method: Functional near-infrared spectroscopy and multiple aspects of personality assessments [the Behavioral Approach System and the Behavioral Inhibition System (BIS) scale, and the Neuroticism–Extroversion–Openness Five–Factor Inventory scale] were applied to explore the neural mechanisms underlying individual differences in deception. Furthermore, a transformer-based model considering personality traits was developed to help people detect deception.

Results: Results showed that personality traits influenced deception primarily through activity in the frontopolar area, dorsolateral prefrontal cortex, and temporoparietal junction area. Moreover, personality traits from different aspects jointly influenced deception, which may vary with behavioral changes. Specifically, extroversion is one of the personality traits partially mediating the effect of BIS on the frontopolar area during deception while being masked during honesty. Additionally, the deception-detection model was found to obtained good classification performance (area under curve: 0.93 ± 0.047), where personality variables contributed to improving the performance.

Conclusions: These findings demonstrated the potential neural impacts of personality in deception for explaining individual differences, which holds great promise in improving the applicability of deception detection for clinical diagnosis.

INTRODUCTION

Although universal, deception may be a symptom of some mental illness, such as antisocial personality disorder (Fullam et al., 2009). In the diagnostic and treatment process that relies on patients' subjective statements and doctors' professional evaluation, patients may conceal their condition or deceive doctors to obtain what they think is good (Rogers and Bender, 2020), especially suicide attempt patients. Approximately 80% of suicide attempt patients do not report suicidal ideation during their last clinical consultation (Gosnell et al., 2019). It is important for healthcare professionals to be aware of this potential deception behavior to provide accurate and effective treatment. Moreover, numerous studies have shown that some specific brain regions, such as the anterior cingulate cortex and inferior frontal gyrus, play a role in guiding deceptive behavior (Speer et al., 2020). Therefore, understanding the activity in these brain regions can help identify deceptive behavior, thus providing valuable insights for clinical diagnosis and treatment. However, individual difference exists in deceptive behavior (Tang et al., 2018), and its underlying neural mechanisms are still not well understood.

Personality comprises relatively stable patterns of behavior, cognition, motivation, and emotional responses (McAdams and Olson, 2010), which is one of the reasons for individual differences in deception (Shao and Lee, 2017). On one hand, personality traits may influence an

* Corresponding author.
E-mail address: huangmeiyan16@163.com (M. Huang).

https://doi.org/10.1016/j.ijchp.2024.100511
Received 6 May 2024; Received in revised form 11 October 2024; Accepted 12 October 2024

individual's tendency to deceive. According to the widely accepted taxonomy of human personality (i.e., the Big Five traits of conscientious, agreeableness, openness, extroversion, and neuroticism) (John and Naumann, 2010), researchers have found that people with high agreeableness engage in minute deceptive behavior (Cuadrado et al., 2021), and that neuroticism is associated with self-deception (Gudjonsson and Sigurdsson, 2004). On the other hand, personality traits may also influence an individual's specific behavioral patterns. Some researchers have explored the role of personality [e.g., the Behavioral Approach System (BAS) and the Behavioral Inhibition System (BIS)] in the potential differences for academic deception based on Gray's reinforcement sensitivity theory (RST) (Bacon et al., 2020). Jung and Lee (2009) also found that different individuals have different implicit preferences for deception decisions in competition tasks. These personality traits constitute an internal inducement system for deceptive behavior, affecting an individual's decision making on whether to perform deceptive behavior, which may lead to differences in brain activation patterns. Extensive research has provided strong behavioral evidence for personality to explain individual differences in deception. However, the exploration of individual differences in the neural mechanisms of deception remains relatively limited. Therefore, exploring the neural impacts of personality traits underlying individual differences in deception can better capture individuals' unique behavioral and brain-activation patterns, which help us detect deception better.

Research shows a close correlation between BAS/BIS and the Big Five traits (Vecchione et al., 2021). For example, high BAS is usually associated with high extroversion and low neuroticism, manifested by more active pursuit of rewards and optimism (Li et al., 2015). Notably, BAS/BIS and Big Five traits are rooted in different research traditions and address different aspects of individual functioning. BAS/BIS primarily relies on biopsychology and neuroscience to explain the mechanisms underlying individual responses to rewards and punishments, emphasizing the role of brain mechanisms and neurotransmitters in regulating behavior and emotional responses (Corr, 2004). Conversely, the Big Five traits represent a descriptive framework for encoding the most salient, socially relevant individual differences by lexical hypothesis (Cattell, 1943). Thus, the BAS/BIS and the Big Five traits, which originate from different theoretical frameworks, influence each other in behavior and emotional responses, providing complementary biological and psychological perspectives. However, most previous studies on the relationship between personality and deception are often based on a single theory of the Big Five traits, rather than exploring it from multiple theoretical perspectives. Given that BAS/BIS is rooted in the basis of the nervous system, numerous studies have also explored the impact of BAS/BIS on related neural mechanisms (e.g., emotional processing, decision making, and motivation) (Bacon, 2020; Li, 2021). Considering the close relationship between BAS/BIS and the Big Five traits, further exploring whether the Big Five traits play a mediating role in the relationship between BAS/BIS and neural activation can help us understand the complex relationship between personality traits and neural mechanisms more comprehensively.

Considering that deception is a complex social activity involving different brain areas, many automatic deception-detection methods, especially systems based on brain-computer interfaces, have been developed and extensively used in brain research. These methods include electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS) (Hu, 2012; Speer, 2020; Wang, 2016). Among them, the fNIRS-based deception-detection system is relatively convenient and cheap, with high temporal resolution compared to fMRI and high spatial resolution compared to EEG. Additionally, most existing deception-detection research focuses on the concealed information test. This passive deception paradigm focuses on the response made by participants in accordance with the examiner's instructions or the experiment cues rather than spontaneous deception in social situations (Farah et al., 2014). Conversely, the active-deception paradigms [e.g., the adapted Ultimatum Game (UG) task] create opportunities for deception in the experiment or interpersonal interaction situations (Güth, 1982; Tang, 2016), which allows participants to choose whether to deceive or not. Therefore, the active deception paradigms can examine more realistic deception behaviors.

Considering the limitations of existing research, the present study aimed to explore the neural mechanisms underlying individual differences in deception through fNIRS and apply it to help people detect deception better. Specifically, we first explored the neural impact of personality during the adapted UG task, where personality traits were assessed from different aspects simultaneously by using the BAS/BIS scale (Li, 2008; Tang, 2018) and the Neuroticism–Extroversion–Openness Five–Factor Inventory (NEO-FFI) scale (Rosellini and Brown, 2011). Whether NEO-FFI mediated the effect of BAS/BIS on neural activation of deception was also investigated. Finally, a potential computer-aided tool was developed using with fNIRS data and personality variables to help people identify deception.

MATERIALS AND METHODS

Participants

Sixty-four college students (age 19–24 years) were recruited for the experiment. Six of them were excluded from the analysis owing to a history of mental illness and loss of fNIRS signal, resulting in a final sample of 58. The participants, who were right-handed with normal or corrected to normal vision, completed a scale comprising a demographic survey and personality assessments (i.e., BAS/BIS and NEO-FFI), followed by the UG task. Specifically, the BAS/BIS scale consists of 20 items that assess an individual's sensitivity to rewards and tendency to avoid punishments, whereas NEO-FFI consists of 60 items designed to comprehensively assess Big Five traits (see detailed item descriptions in Supplementary Materials). The descriptive statistics, reliabilities, and inter correlations among personality traits are presented in Table 1. All personality assessments had good internal consistency (Cronbach's $\alpha >$ 0.60).

Written informed consent was obtained from all participants. This study was approved by the Ethical Review Board of Southern Medical University and conducted according to the Declaration of Helsinki.

Task and procedures

During the experiment, participants performed a scale consisting of a demographic survey and personality assessments (i.e., BAS/BIS and NEO-FFI), followed by an adaptation of the UG task. This task involved a proposer and a responder to allocate a sum of money in complete anonymity, where the proposer suggested an allocation scheme and the responder had the right to choose to accept or reject it. In particular, the money was distributed according to plan if the responder accepted it; otherwise, both parties lost out. During this process, the responder was unaware of the true total amount of money distributed, so that the proposer can opt to tell the responder the true total amount or a false amount for her own benefit. Therefore, whether participants deceive others for their own benefit (i.e., honest or deceitful decisions) can be investigated by observing the total amount they reported. In fact, deception is an intentional behavior, whose essence is to mislead others by conveying false beliefs (Ding, 2014; Saini, 2021). This definition means that the truth in the interactive scenario may also carry deceptive intent. To exclude the interference of this factor, a control task in which participants performed the same adapted UG task was established. In the control task, the proposer could only disclose the true total amount to the receiver (i.e., honest decision). Then, the neural-activation comparison of the honest decision phase between the two tasks was used to infer whether the honest decision of the UG task has deception intent or not. Each task contained 20 trials. The general experimental procedure is shown in Fig. 1A. In this study, all participants were proposers, and

Table 1
Descriptive statistics, reliabilities, and inter correlations among variables.

	Mean	SD	1	2	3	4	5	6	7
1.Conscientious[v]	41.66	5.74	**0.83**						
2.Agreeableness[v]	41.15	4.90	0.44**	**0.63**					
3.Openness[v]	42.94	5.43	–0.01	–0.31*	**0.71**				
4.Extroversion[v]	36.83	6.31	0.32*	0.23	0.16	**0.81**			
5.Neuroticism[v]	34.62	8.22	0.55**	–0.39**	0.34*	–0.21	**0.89**		
6.BAS[v]	38.34	3.31	0.45**	–0.04	0.35*	0.31*	–0.21	**0.68**	
7.BIS[u]	14.55	2.14	–0.37*	–0.11	–0.07	–0.28*	0.49**	–0.08	**0.70**

[v] indicates that the variable is normally distributed and the Pearson correlation analysis is used.
[u] indicates that the variable is not normally distributed and the Spearman correlation analysis is used.
$p_{FDR}{}^{*} < 0.05$ *and* $p_{FDR}{}^{*} < 0.01$. Reliability coefficients are shown in bold along the diagonal of the table.

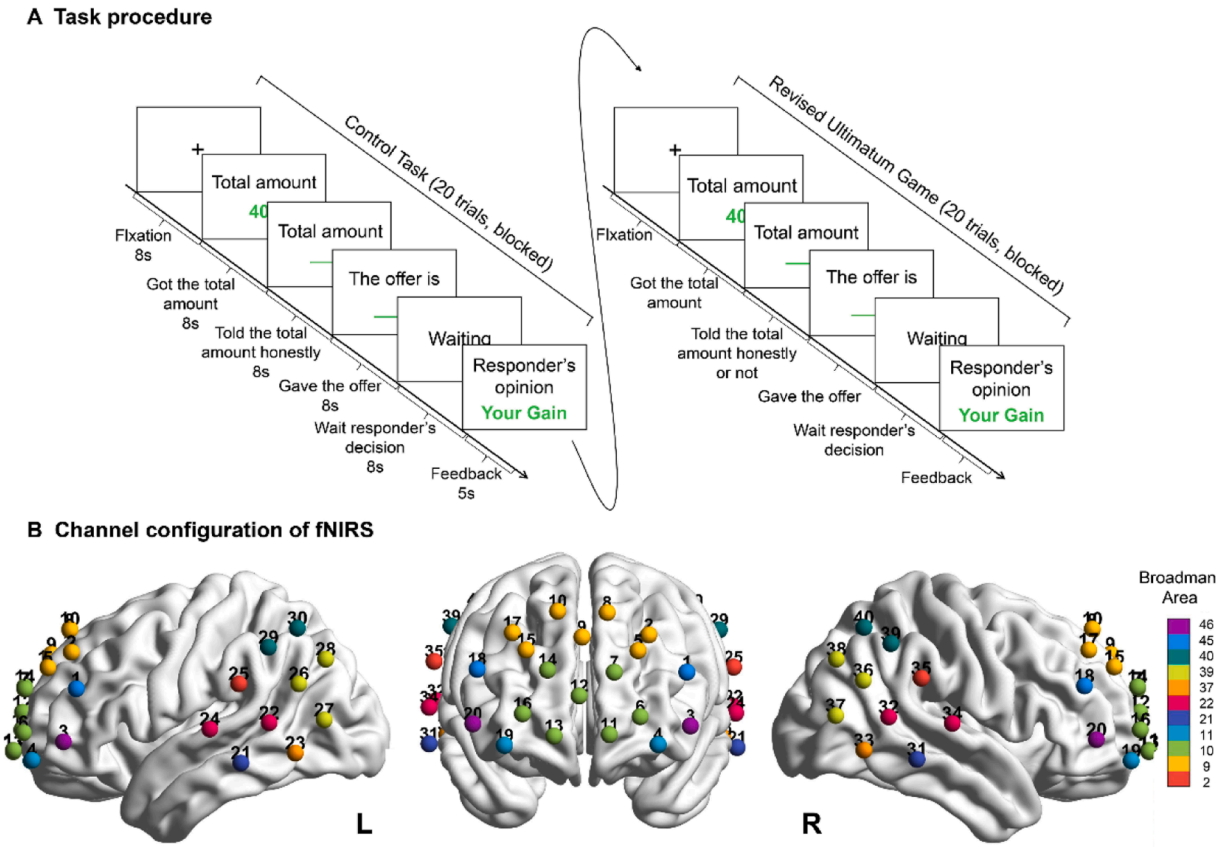

Fig. 1. Task and fNIRS setup of the experiment. (A) During this experiment, participants, acting as the proposer, perform the control task and the adapted UG task. In the control task, participants can only honestly tell the responder the total amount; in the UG task, they can opt to say the true total amount or not according to their own benefits. Each task consists of 20 trials. (B) 40 channels of fNIRS are mainly covered in the prefrontal, temporal, and parietal cortices, where the different colors represent the coverage for the corresponding Brodmann brain regions.

respondents were computer-generated characters. The responders' acceptance or rejection was randomly arranged and had no bearing on the proposer's actual allocation. Details of the task were disclosed to the participants at the end (see detailed behavioral results in Supplementary Materials).

Acquisition of fNIRS data

During the experiment, the NIRScout system (NIRx, USA) was used to collect imaging data from participants and measure the absorption of near-infrared light at two wavelengths (785 and 830 nm). According to the international 10–20 system, the probe placement comprising 40 measurement channels (15 dual-wavelength laser diodes and 16 detectors), covered the prefrontal, temporal, and parietal cortices, which were then registered to the Brodmann area (Fig. 1B). The color bar shows the probe placement covered the Brodmann area with 2, 9, 10, 11, 21, 22, 37, 39, 40, 45, and 46. For each participant, the probe set was examined and adjusted to ensure consistency in the position. Changes in oxyhemoglobin (HbO) and deoxyhemoglobin (HbR) concentrations in each channel were recorded according to the modified Beer–Lambert law. This study primarily focused on the changes in HbO concentration, which have been proven to be the most sensitive indicator of regional cerebral blood-flow changes in fNIRS measurements.

Analysis between brain and personality traits

To explore the neural impact of personality on deception, neural-activation analysis and functional-connectivity analysis were given primarily focus. The acquired fNIRS data were primarily pre-processed and analyzed using the nirsLab software accompanying the NIRScout system. First, data from 15 s before the task started to 10 s after the task ended were retained for further analysis, whereas other unneeded data

were removed. Baseline correction and removal of spike artifacts followed. The data were then band-pass filtered at 0.01–0.2 Hz to remove the noise. Then, the hemodynamic response was calculated according to the modified Beer–Lambert law. Subsequently, the *beta* values of the HbO concentrations were calculated for each channel by the generalized linear model, and further differences in *beta* values between activated channels were examined by *t*-test. Finally, functional connectivity was calculated based on the mean fNIRS data by Pearson correlation.

To ensure that the honesty decision of the UG task was unaffected by deception intention, the neural activation of honesty decision phase in the control task was first compared with that in the UG task. Then, Pearson correlation analyses were performed between personality and the *beta* values of neural activation in different phases of the UG task (i.e., preparation, honest/deceitful decision, and feedback).

To further explore how different brain regions interacted during deceitful or honest decisions, we investigated the correlations of personality traits and functional connectivity between brain regions during the decision phase of the UG task. After obtaining a correlation matrix of honest and deceitful decisions, Pearson correlation analysis was performed on functional connectivity and personality-trait scores. Finally, all resulting empirical *p* values were then corrected for multiple comparisons through the false discovery rate (FDR) method.

Mediation analysis

Mediation analysis was conducted to understand the role of different personality traits further in deception. Given that RST (i.e., BAS/BIS) has a strong biological basis, the mediating effect of NEO-FFI in the relationship between BAS/BIS and neural activation was examined. The above analyses were performed by applying the mediation model in the SPSS macro developed by (Hayes and Preacher, 2014), and the percentile Bootstrap method was used to test the significance of the mediation effect.

Deception detection based on fNIRS data

We attempted to understand the neural mechanisms involved in individual difference of deception and strived to develop a computer-aided tool to help people identify deception. Referring to Wang et al. (2022), a transformer-based neural network was used to distinguish between deception and honesty. On this basis, we further incorporated personality variables for deception detection (Fig. 2).

The spatial level and channel level are different approaches to improving data utilization. Accordingly, a two-branch model structure was used. Specifically, one-layer convolution operations with different kernel sizes were first performed on fNIRS data $x \in R^{H\times W\times C}$ respectively, where H is the number of the channel, W is the number of time point, and C is the dimension of hemoglobin (i.e., HbO and HbR). In the first branch, the spatial-level convolution (i.e., $\mathrm{Conv_S}$, kernel size: 5×30) concentrated on different activation regions of multiple channels. In the second branch, the channel-level convolution (i.e., $\mathrm{Conv_C}$, the kernel size: 1×30) focused on the cerebral hemodynamic response of a single channel. Then, the features from the two branches were rearranged along the W axis to maintain fNIRS channel-location information, further dividing into H_S sequences and H_C sequences followed by linear embedding and layer normalization.

$$x_S = \mathrm{embedding}(\mathrm{Conv_s}(x)) \tag{1}$$

$$x_C = \mathrm{embedding}(\mathrm{Conv_C}(x)) \tag{2}$$

Subsequently, the two sequences (i.e., x_S and x_C) were fed into the transformer encoder **E** to assign different attention weights to brain activation and hemodynamic responses. The transformer encoder **E** comprised six identical layers, each with two main sub-layers: a multi-head self-attention and a multi-layer perceptron (MLP). Specific structural information of the transformer encoder can be found in Vaswani et al. (2017). Then, the features obtained from **E** were fused with personality variables x_{per} through a concatenation operation. Finally, the fused features were input into a three-layer MLP module, where the node number of each layer was 135, 64, and 2, respectively, to obtain the predicted results $\hat{y}$.

$$\hat{y} = \mathrm{MLP}\{\mathrm{concat}[\mathbf{E}(x_S), \mathbf{E}(x_C), x_{pv}]\} \tag{3}$$

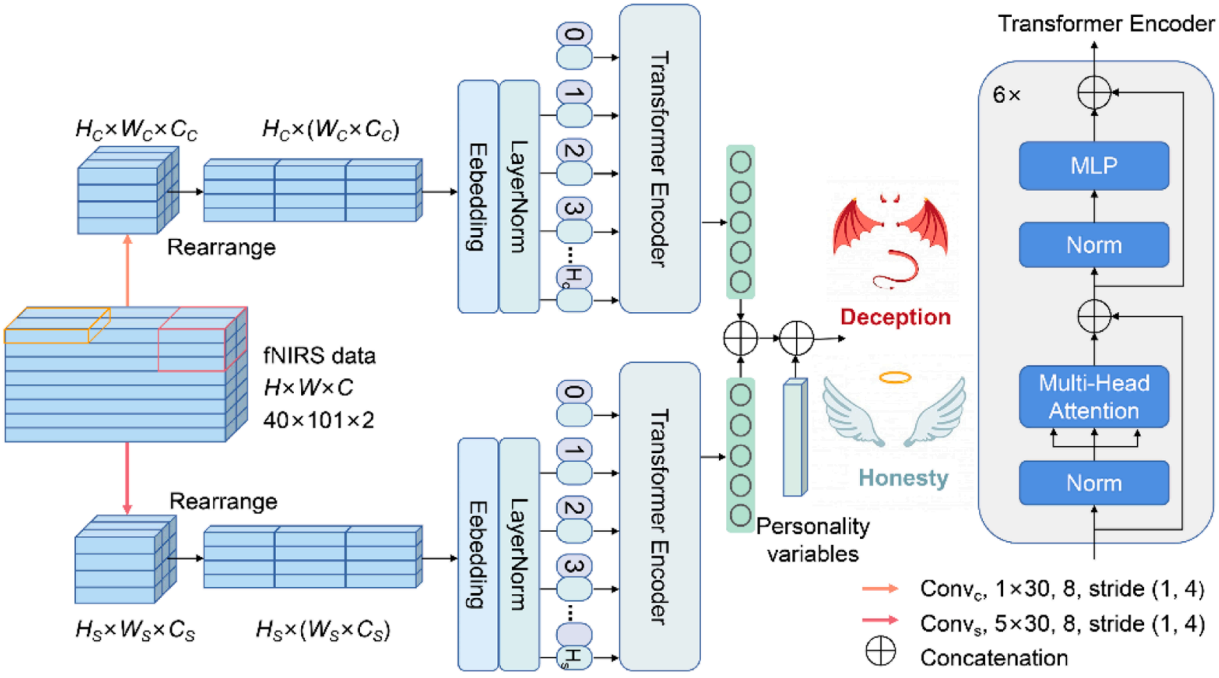

Fig. 2. Overview of the automatic deception-detection method. The proposed method is composed of spatial- (down) and channel-branch (up). In the two branches, convolution operations with different kernel sizes were performed on fNIRS data, which concentrates on different activation regions of multiple channels and the cerebral hemodynamic response of a single channel, respectively.

Cross-entropy loss function is applied as the loss function of the proposed network, which can be defined as

$$\ell = \sum_{k=0}^{1} -y_k \log(\widehat{y}_k) \tag{4}$$

where y is 1 for the deception class and 0 for the honesty class, and y is the probability prediction of each label $k \in \{0, 1\}$. To avoid overfitting and overestimation, some strategies (i.e., label smoothing and flooding) are introduced for model training.

For label smoothing (Szegedy et al., 2016), y is smoothed by the soothing parameter α that is usually set to 0.1:

$$y_k^{ls} = (1-\alpha)y_k + \frac{\alpha}{2} \tag{5}$$

For flooding (Ishida et al., 2020), it can keep the training loss around a small value called flooding level to avoid zero training loss and reduce the test loss. Thus, the final loss function is defined as

$$\widetilde{\ell} = \left| \sum_{k=0}^{1} -y_k^{ls} \log(\widehat{y}_k) - b \right| + b \tag{6}$$

where b is the flooding level. The model is trained for 120 epochs with a batch size of 128 by using AdamW with an initial learning rate of 0.001. In addition, flood levels are attenuated over specific periods to obtain a continuous regularization effect (i.e., the flood levels over specific periods in the network (epoch, flooding level) are (1, 0.30), (20, 0.40), and (50, 0.30).

As shown in Fig. 1A, each participant underwent 20 trials in each task, where each trial lasted 32 s. The fNIRS data were divided into independent samples as input in accordance with trial duration, whereas the label was determined by participants' decision in the current trial. Some participants made deception decisions throughout the UG task, and thus, some participants' samples had only positive cases (i.e., deception). Accordingly, the last 10 trials of the control task were also included as honest samples to avoid this situation. Finally, a dataset with a sample size of 1740 were obtained, including 742 cases of deception and 998 cases of honesty. To avoid overfitting and overestimation, a hold-out strategy in accordance with participants was applied for the evaluation of models. All participants were divided into 10 subsets: eight were applied as the training set to train the networks, one was used as the validation set to guide the halting of network optimization and parameter tuning, and one was used as the testing set to evaluate the network. For example, six participants' data were utilized as the testing set and six as the validation set; the rest were in the training set, so the sample sizes of the testing and validation sets were 180 ($30 \times 6 = 180$), and the sample sizes of the training set were 1380 ($30 \times (58 - 6 \times 2) = 1380$). This training was repeated five times, and the final result was obtained by averaging the test results. In this study, area under the curve (AUC), accuracy (ACC), specificity, sensitivity, and F1 score were utilized as evaluation indicators for the model, whose value closer to 1 is indicative of superior performance (see details in the Supplementary Materials).

The gradient-weighted class activation mapping (Grad-CAM) of the channel-level branch in the trained model was calculated to investigate the brain regions of interest to the model. Moreover, a word-cloud analysis with meta-analysis derived studies obtained from Neurosynth (https://www.neurosynth.org/) was conducted on channels of interest based on the results of Grad-CAM (threshold = 0.70).

RESULTS

Personality associated with neural activation during the UG task

Considering that truth in interactive situations may also carry deceptive intention, the neural activation of the honest decision phase in the control task was first compared with that in the UG task. Results showed no significant difference at both tasks, so we assumed that truth in the UG task did not carry deceptive intention. Subsequently, the relationship between personality and neural activation in various phases of the UG task was investigated.

After FDR correction, personality traits were significantly associated with the activation of different brain regions ($p_{FDR} < 0.05$). During the preparation phase of the UG task (i.e., obtaining the total amount), BAS and BIS were significantly correlated with neural activation among channels primarily covering the dorsolateral prefrontal cortex (DLPFC), orbitofrontal area (OFC), pars triangularis Broca's area, superior temporal gyrus (STG), and temporoparietal junction area (TPJ). By contrast, only the specific dimension of personality description (i.e., openness) was significantly associated with neural activation among two channels covering DLPFC and frontopolar area (Fig. 3A). These findings indicated that BAS- and BIS-sensitive individuals may be more likely to regulate current activities according to their own purposes. During the honesty phase of the UG task, personality traits presented a positive correlation trend with neural activation, where the correlations among neuroticism and STG, agreeableness and DLPFC as well as STG, and extroversion and fusiform gyrus were significant (Fig. 3B). Conversely, personality traits except agreeableness showed a negative correlation trend with neural activation during the deception phase of the UG task, where significant correlations were observed, including agreeableness and TPJ, openness and TPJ as well as STG, BIS and OFC as well as TPJ, and BAS and OFC (Fig. 3C). This result may be related to the fact that in comparison to honesty, more attention was required to suppress the associated brain activity during deception. During the feedback phase of the UG task, personality traits except extroversion and conscientious presented a negative correlation with the neural activation of prefrontal cortex (Fig. 3D), which was sensitive to rewards (Zhang and Gläscher, 2020). The above findings suggested that BAS/BIS and NEO-FFI had different effects on neural activation in different situations from different perspectives.

Relationships between personality and functional connectivity of deception

To explore how different brain regions interacted in the decision to be deceitful or honest, the correlation between personality traits and functional connectivity during the decision phase of the UG task was further analyzed. Correlation analyses showed that personality traits were significantly correlated with functional connectivity only in the deception phase of the UG task after FDR correction ($p_{FDR} < 0.05$). Specifically, openness was significantly correlated with connections (Fig. 4A) in ch23–ch40 (middle temporal gyrus and TPJ), and BIS was significantly correlated with connections (Fig. 4B) in ch7–ch8 (frontopolar area and DLPFC), ch7–ch25 (frontopolar area and primary somatosensory cortex), ch13–ch27 (frontopolar area and TPJ), and ch23–ch27 (fusiform gyrus and TPJ). The experimental results revealed that BIS-sensitive individuals were more relevant in the left hemisphere of neural connections (e.g., frontopolar area and TPJ).

Mediation analysis

To understand the role of personality traits from different aspects in deception, we further explored whether NEO-FFI was mediating the effect of BAS/BIS on brain during the decision phase of the UG task. Table 2 shows the significant mediation effects of personality in NEO-FFI, mediating the effect of BAS/BIS on neural activation. During the honesty phase of the UG task, extroversion had a masking mediation effect between BIS and the neural activation of the frontopolar area. During the deception phase of the UG task, extroversion mediated the causal effect of BAS on the neural activation of DLPFC, and openness mediated on fusiform gyrus, which was the masking effect. Conversely, extroversion played a partial mediating effect in BIS and the frontopolar area. Notably, extroversion had different mediation effects between BIS

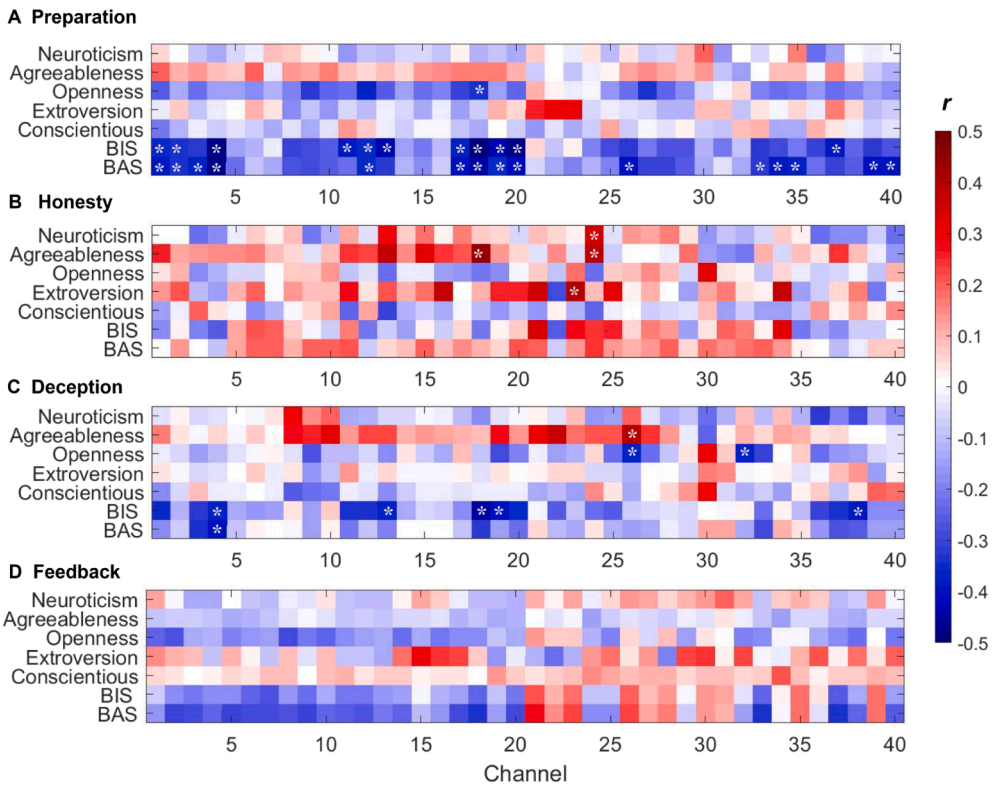

Fig. 4. Relationships between personality and functional connectivity during the deception phase of UG task. (A) Openness is significantly correlated with connections in middle temporal gyrus and TPJ. (B) BIS is significantly correlated with connections in frontopolar area, DLPFC, primary somatosensory cortex, TPJ, and fusiform gyrus.

and the neural activation of the frontopolar area in different contexts, with a masking effect during the honesty phase and a partial mediating effect during the deception phase. This finding indicated that the mediating effect NEO-FFI on BAS/BIS and neural activation varied with situational changes (Fig. 5A). The above evidence demonstrated that NEO-FFI and BAS/BIS can describe personality from different perspectives and have different effects on brain activities.

Deception detection based on fNIRS data

To evaluate the performance of the model, the model was compared with support vector machine (SVM), spatial-level transformer, channle-level transformer, and two-branch convolutional neural network (CNN) methods. Among them, the dual branch CNN is similar to the two-branch transformer architecture, where the transformer encoder is replaced by the corresponding convolution layer. The comparison methods utilizing deep learning employed same input and training strategies as the proposed model. For the SVM, the inputs consisted of the personality variable and the statistical features (i.e., average, peak, slope, kurtosis, maximum, and minimum) of hemoglobin (i.e., HbO and HbR) from each channel of the samples (Huang et al., 2022). As shown in Table 3, the performance of the two-branch transformer model is superior to other models, which indicates that transformer has great potential for the exploitation of fNIRS features. In addition, the performance of the two-branch model is better than that of spatial-level and channel-level models, which indicates that fNIRS information can be mined more effectively by combining spatial and temporal characteristics. Fig. 5B shows the performance of the deception-detection model with/without personality variables. Results confirmed that model performance improved to a certain extent after adding personality variables. Among them, significance existed in ACC ($p < 0.05$). This finding laterally indicated that personality traits can explain individual differences in the deceptive brain activity and show potential to improve performance in deception detection. The Grad-CAM of the model was further calculated to investigate the brain regions of interest to the model (Fig. 5C). The model without personality variables primarily focused on OFC and TPJ, whereas the model with personality variables paid more attention to the frontopolar area and DLPFC. In other words, the model considered that personality variables may go through the frontopolar area and DLPFC to influence deceptive brain activity, which was generally consistent with the above findings. Furthermore, a word-cloud analysis with meta-analytically derived studies obtained from Neurosynth showed that these brain regions were mostly related to memory, cognitive control, emotion, visual processing, and social cognition (Fig. 5D), whose cognitive functions are related to the processing mechanisms of deception in different situations.

DISCUSSION

Personality is one of the known explanations for individual differences in deceptive behavior, but the underlying neural mechanisms remain unclear. In this study, we used fNIRS and multiple aspects of personality assessment (BAS/BIS and NEO-FFI) to explore the neural mechanisms underlying individual differences in deception. Furthermore, we established a model based on personality traits and fNIRS to help people identify deception, thereby advancing our understanding of deception. The results are discussed below.

First, the influence of personality traits in neural activation as well as functional connectivity was explored to explain individual differences in deception. On the one hand, the neural activation results showed that different personality traits influenced neural activation during different phases of the UG task. Specifically, most personality traits were positively correlated with most neural activation (e.g., prefrontal cortex, STG, and TPJ) when honest but negatively correlated when deceitful, whereas agreeableness showed a positive correlation regardless of neural activation when deceitful (e.g., TPJ) or honest (e.g., pars

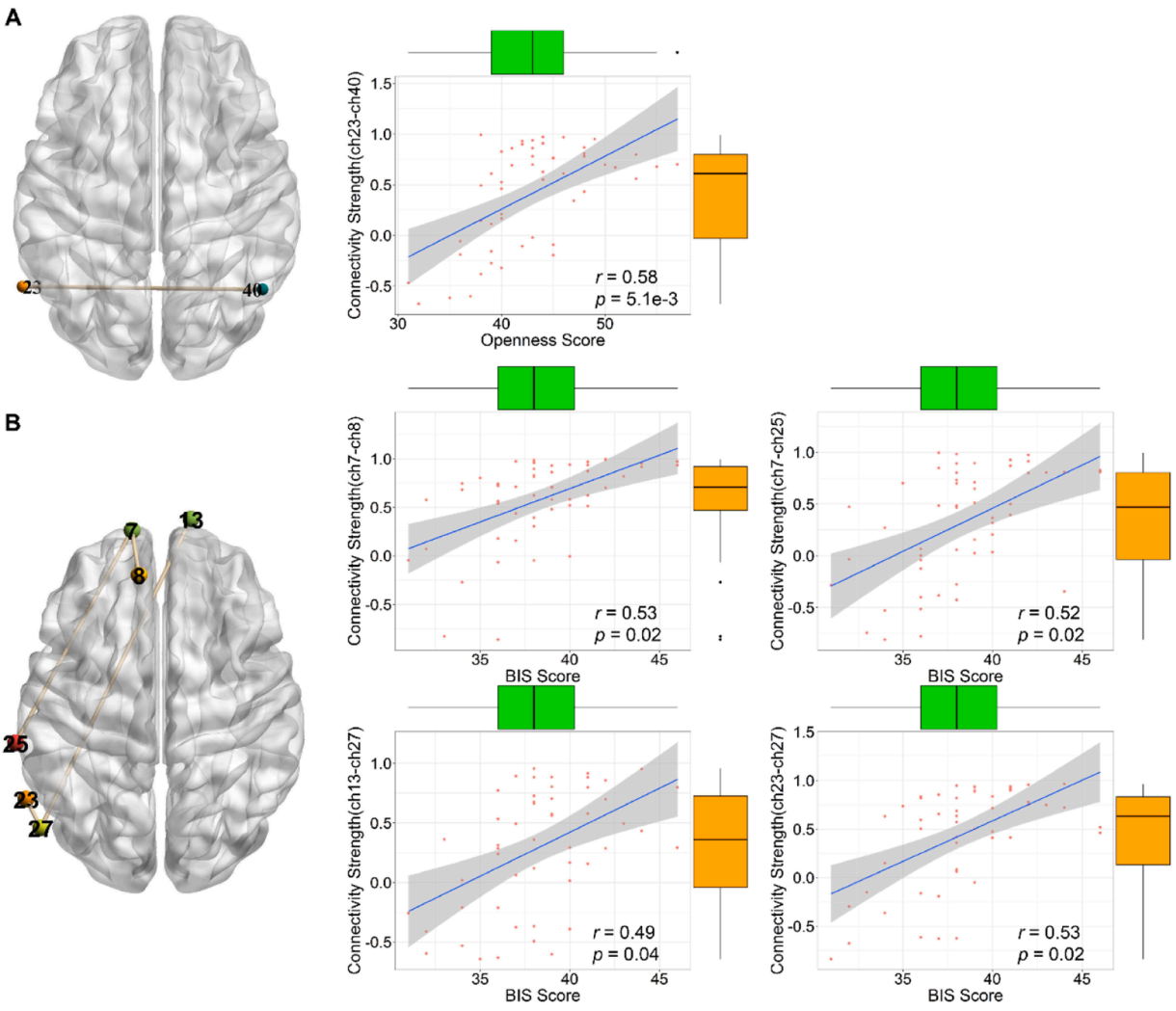

Fig. 3. Association between brain activity and personality traits during UG task. (A), (B), (C), and (D) show the association of neural activation and personality traits during the preparation, honesty, deception, and feedback phases of the UG task, respectively. * indicates the association is significant after FDR correction ($p_{FDR} < 0.05$).

Table 2
Results of bootstrap analysis of mediating effect.

Condition	Path	Total effect	Direct effect	Indirect effect	Boot LLCI	Boot ULCI
Honesty	BIS→Extroversion→ch14	–0.32	0.25	–0.07	–0.15	–0.01
Deception	BAS→Extroversion→ch9	0.22	0.18	0.04	1.9e-3	0.08
	BAS→Openness→ch27	0.16	0.24	–0.08	–0.20	–1.3e–3
	BIS→Extroversion→ch11	–0.58	–0.48	–0.10	–0.24	–3.0e–3

triangularis Broca' s area and STG). Among the neural activations, the prefrontal cortex including the DLPFC, is primarily associated with response inhibition (Abe, 2011); TPJ is involved in theory of mind (i.e., the ability to infer the mental states of others) (Bhatt et al., 2010); and pars triangularis Broca' s area and STG are jointly responsible for language understanding (Engelmann, 2019; Grosse Wiesmann, 2020). These findings suggested that most personality traits (e.g., openness, BAS, and BIS) required response-inhibition activities to maintain specific cognitive control (e.g., inhibit the true response) or attentional focus on the other (e.g., infer the other's beliefs) during deception. Agreeableness had different neural activation patterns. As a result of these neural activation, individuals with high agreeableness may understand the mental states of others easily and pay more attention to verbal communication when faced with deceptive situations, thereby affecting their behavioral performance on deception. This may be related to the fact that individuals with high agreeableness tend to maintain harmonious interpersonal relationships and are willing to compromise and cooperate to resolve conflicts.

On the other hand, the relationship between functional connectivity and personality was significant only during deception. Openness scores were significantly positively correlated with the connectivity between the left fusiform gyrus and the right TPJ. The fusiform gyrus is involved in object recognition (e.g., impressions and words) (Gerrits, 2019; Spagna, 2021). According to previous studies, individuals need to model their image continuously in the minds of their opponents when deceiving to perform impression management for building trust (Luo et al., 2017). Individuals with high openness, who may think about problems from multiple perspectives, focus on impression management

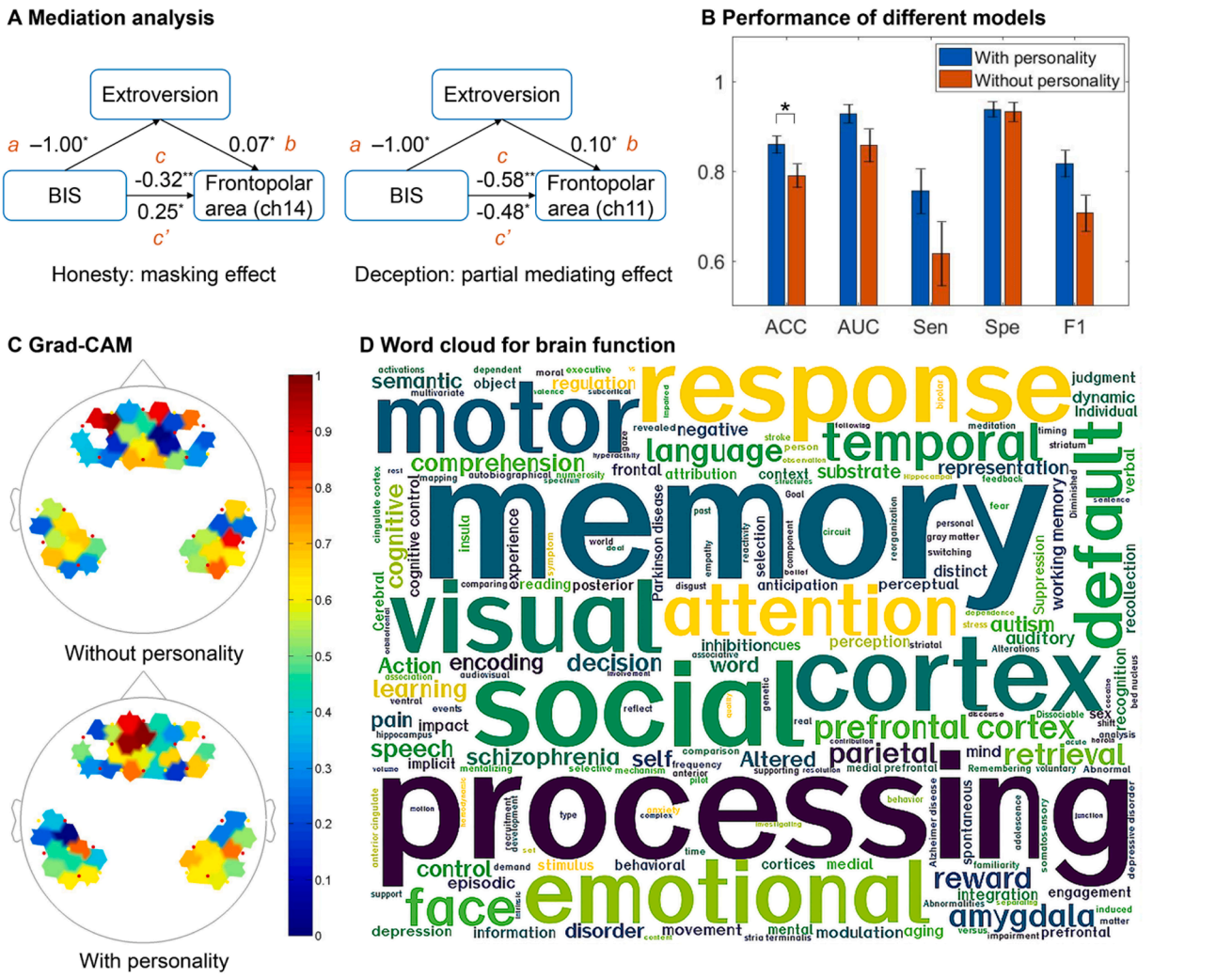

Fig. 5. Results of mediation analysis and classification. (A) Extroversion had different mediation effects between BIS and neural activation of the frontopolar area with situational changes. (B) The performance of the model with/without personality variables was accessed by accuracy (ACC), area under curve (AUC), specificity (Spe), sensitivity (Sen), and F1 score, and * indicates significance difference ($p < 0.05$). (C) The Grad-CAM of each channel in the transformer model was calculated, where the closer the color is to red, the more attention the model pays to the channel. (D) The word cloud analysis of the channels that the model with personality focused on was conducted by meta-analytically derived studies from Neurosynth.

Table 3
Performance comparison of different models.

	AUC	ACC	Specificity	Sensitivity	F1-score
SVM	0.68 ± 0.063	0.65 ± 0.103	0.75 ± 0.089	0.60 ± 0.143	0.62 ± 0.078
Dual branch CNN	0.92 ± 0.047	0.80 ± 0.084	0.86 ± 0.072	0.75 ± 0.213	0.75 ± 0.133
Spatial-level transformer	0.85 ± 0.082	0.77 ± 0.057	0.87 ± 0.100	0.63 ± 0.124	0.70 ± 0.089
Channel-level transformer	0.88 ± 0.077	0.80 ± 0.068	0.91 ± 0.070	0.67 ± 0.125	0.74 ± 0.095
Dual branch transformer	**0.93 ± 0.047**	**0.86 ± 0.043**	**0.94 ± 0.039**	**0.76 ± 0.111**	**0.82 ± 0.065**

through the functional connectivity between the left fusiform gyrus and the right TPJ during deception. Moreover, BIS was correlated with the strength of functional connectivity in the left hemisphere with the lateralization function of seeking explanations for internal and external events, and they are sensitive to punishment signals and negative emotions (Sommer et al., 2016). Thus, BIS may drive individuals to rationalize their deceptive behavior through the above connectivity. These findings contributed to the current understanding of how different personality traits affected the neural mechanisms of deception.

Second, previous deception studies focusing on individual differences use only personality assessments derived from neurobiological approaches (e.g., BAS/BIS) or from the lexical hypothesis (e.g., NEO-FFI). Accordingly, the current work explored the role of personality assessments from these two aspects in explaining individual differences of deception. Consistent with previous findings, NEO-FFI and BAS/BIS may help describe individuals in a relatively consistent pattern of behavior. Results of mediation analysis revealed that openness played a masking effect in the influence of BAS on fusiform gyrus activation during deception. In line with previous findings, openness was observed to be associated with impression management during deception. Conversely, BAS, the root of impulsive personality, is likely to ignore impression management during deception. Therefore, openness played a negative role in the relationship between BAS and the activation of the fusiform gyrus. Extroversion also exerted a partial mediating effect on the influence of BAS in DLPFC and on the influence of BIS in the frontopolar area. Conversely, extroversion played a masking effect on the influence of BIS in the frontopolar area when honest. DLPFC played an important role in response inhibition, and frontopolar area contributed to inferring others' beliefs or mental states (Abe, 2011). Thus, extroversion may facilitate the ability of BAS-sensitive individuals to inhibit

original responses (i.e., honesty) and achieve goal-directed behavior, which may result from extroversion and BAS being closely related to the tendency to seek stimulation and reward (Sommer et al., 2016). Extroversion also facilitated BIS-sensitive individuals' inferences about other people's intentions when they were deceiving but inhibited their inferences when they were honest. Given that BIS individuals tend to exhibit higher anxiety and cautious behaviors, extroversion may facilitate BIS-sensitive individuals' inferences about other people's intentions when deceiving by enhancing their social skills and theory-of-mind reasoning ability. When being honest, it prompted them to focus more on their current performance rather than inferring others' intentions. Taken together, the above results indicated that the mediating effect of NEO-FFI on BAS/BIS and neural activation may change with individual behavioral changes, which may help elucidate the relationship between personality traits and neural functions more comprehensively.

Third, we further developed a transformer-based model to help people identify deception. Compared to traditional machine learning methods, deep learning has become a popular choice for analyzing and processing medical images (Wang, 2018, 2019; Zhao, 2023). In particular, transformers have strong learning capabilities for capturing long-range interactions (Wang et al., 2024), which shows excellent performance in mining fNIRS data. Additionally, the inclusion of personality variables can improve the model's performance, as personality variables help the model better capture individual unique patterns of neural activation. Consistent with the above results, the model with personality variables focused more on the frontopolar area and DLPFC, whereas the model with personality variables focused primarily on OFC and TPJ. Briefly, the deep learning model with automatic learning suggested that personality variables may modulate deceptive brain activity through the prefrontal cortex, which is closely related to the process of response inhibition (Abe, 2011). During deception, truth-telling is often a dominant response and a spontaneous behavior for individuals, whereas deception requires the suppression of this dominant response. Furthermore, word clouds generated based on Neurosynth meta-analysis also revealed that most of these brain regions were associated with memory, cognitive control, emotion, visual processing, and social cognition, which are involved in deceptive processing mechanisms in different situations. Accordingly, deep learning provided a new perspective to explain that individuals with different personalities performed deception by influencing individuals' decision making and inhibiting the dominant response of truth telling.

In this study, there are several limitations. First, the relatively small sample size limits the further comparative research on the neural mechanisms of deception in different personality traits, such as the grouping of high agreeableness and low agreeableness, and limits the further development of deception detection models. Future work will need to replicate and extend these findings in a larger sample of individuals. Second, the sample population in this study lacked diversity, which may limit the broader applicability of our conclusions. Future research should recruit participants from diverse backgrounds in terms of age, gender, cultural context, and psychological characteristics to explore how these factors interact with deception and related neural mechanisms. Finally, the responses of respondents in UG tasks are automatically generated by a computer program, which limits the ecological benefits to some extent. In the future, we can collect two-person interpersonal neural synchronization data in the simulated situation of clinical diagnosis, which more realistically reflects the actual situation of deception.

Collectively, this study offered support for a biological theory concerning the neural effects of personality on deception. It also applied this theory to enhance the reliability and applicability of deception detection. The neural impacts involving the frontopolar area, DLPFC, and TPJ appeared to be the neural basis for the personality interpretation of individual differences in deception, which are primarily related to response inhibition, social cognition, and theory of mind. Moreover, multiple theoretical perspectives of personality assessment were performed in this study and revealed that the mediating effect of personality assessed by NEO-FFI on BAS/BIS and neural activation may vary with individual behavioral changes. Finally, the transformer-based deception-detection model obtained good classification performance and took a new perspective on personality to explain the individual differences in the neural mechanism of deception.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 82272069, No. 81601562, No. 81974275, and No. 12126603); the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515012023); and the Science and Technology Planning Project of Guangzhou (No. 201904010417).

Data and code availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. The code of the automatic deception-detection method is available at https://github.com/Meiyan88/fNIRS-deception-detection.

References

Abe, N (2011). How the brain shapes deception: An integrated review of the literature. *Neuroscientist., 17*, 560–574.
Bacon, AM, et al. (2020). What motivates academic dishonesty in students? A reinforcement sensitivity theory explanation. *Br J Educ Psychol, 90*, 152–166.
Bhatt, MA, et al. (2010). Neural signatures of strategic types in a two-person bargaining game. *Proc Natl Acad Sci U S A., 107*, 19720–19725.
Cattell, RB (1943). The description of personality: Basic traits resolved into clusters. *J Abnormal and Social Psychology, 38*, 476.
Corr, PJ (2004). Reinforcement sensitivity theory and personality. *Neurosci Biobehavl Rev, 28*, 317–332.
Cuadrado, D, et al. (2021). Personality, intelligence, and counterproductive academic behaviors: A meta-analysis. *J Pers Soc Psychol, 120*, 504.
Ding, XP, et al. (2014). Neural correlates of second-order verbal deception: A functional near-infrared spectroscopy (fNIRS) study. *NeuroImage, 87*, 505–514.
Engelmann, JB, et al. (2019). The neural circuitry of affect-induced distortions of trust. *Sci Adv, 5*, eaau3413.
Farah, MJ, et al. (2014). Functional MRI-based lie detection: scientific and societal challenges. *Nat Rev Neurosci, 15*, 123–131.
Fullam, RS, et al. (2009). Psychopathic traits and deception: functional magnetic resonance imaging study. *Br J Psychiatry, 194*, 229–235.
Gerrits, R, et al. (2019). Laterality for recognizing written words and faces in the fusiform gyrus covaries with language dominance. *Cortex, 117*, 196–204.
Gosnell, SN, et al. (2019). Classifying suicidal behavior with resting-state functional connectivity and structural neuroimaging. *Acta Psychiatr Scand, 140*, 20–29.
Grosse Wiesmann, C, et al. (2020). Two systems for thinking about others' thoughts in the developing brain. *Proc Natl Acad Sci U S A., 117*, 6928–6935.
Gudjonsson, GH, & Sigurdsson, JF (2004). The relationship of suggestibility and compliance with self-deception and other-deception. *Psychol Crime Law, 10*, 447–453.
Güth, W, et al. (1982). An experimental analysis of ultimatum bargaining. *J Econ Behav Organ, 3*, 367–388.
Hayes, AF, & Preacher, KJ (2014). Statistical mediation analysis with a multicategorical independent variable. *Br J Math Stat Psychol, 67*, 451–470.
Hu, X-S, et al. (2012). fNIRS-based online deception decoding. *J Neural Eng, 9*, 026012.
Huang, M, et al. (2022). Joint-channel-connectivity-based feature selection and classification on fNIRS for stress detection in decision-making. *IEEE Trans Neural Syst Rehab Eng, 30*, 1858–1869.
Ishida T, et al. (2020), Do we need zero training loss after achieving zero training error? arXiv preprint arXiv: 200208709.
John, OP, & Naumann, LP (2010). Surviving two critiques by block? The resilient big five have emerged as the paradigm for personality trait psychology. *Psychol Inq, 21*, 44–49.
Jung, KH, & Lee, J-H (2009). Implicit and explicit attitude dissociation in spontaneous deceptive behavior. *Acta Psychologica, 132*, 62–67.
Li, C-W, et al. (2021). Motivational system modulates brain responses during exploratory decision-making. *Sci. Rep., 11*, 15810.

Li, Y, et al. (2015). Big-five personality and BIS/BAS traits as predictors of career exploration: The mediation role of career adaptability. *J Vocat Behav, 89*, 39–45.
Li, Y, et al. (2008). The Chinese version of the BIS/BAS scale: reliability and validity. *Chinese Mental Health Journal, 22*, 613–616.
Luo, Q, et al. (2017). The functional architecture of the brain underlies strategic deception in impression management. *Front. Hum. Neurosci., 11*, 513.
McAdams, DP, & Olson, BD (2010). Personality development: Continuity and change over the life course. *Annu. Rev. Psychol., 61*, 517–542.
Rogers, R, & Bender, SD (2020). *Clinical assessment of malingering and deception*. New York: Guilford Publications.
Rosellini, AJ, & Brown, TA (2011). The NEO Five-Factor Inventory: Latent structure and relationships with dimensions of anxiety and depressive disorders in a large clinical sample. *Assessment, 18*, 27–38.
Saini, N, et al. (2021). Information detection in brain using wavelet features and K-nearest neighbor. In *2021 6th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1704–1709). IEEE.
Shao, R, & Lee, TMC (2017). Are individuals with higher psychopathic traits better learners at lying? Behavioural and neural evidence. *Transl Psychiatry, 7*. e1175-e1175.
Sommer, K, et al. (2016). BIS/BAS sensitivity and emotional modulation in a prepulse-inhibition paradigm: A brain potential study. *Physiol. Behav., 154*, 100–113.
Spagna, A, et al. (2021). Visual mental imagery engages the left fusiform gyrus, but not the early visual cortex: A meta-analysis of neuroimaging evidence. *Neurosci Biobehav Rev, 122*, 201–217.
Speer, SPH, et al. (2020). Cognitive control increases honesty in cheaters but cheating in those who are honest. *Proceedings of the National Academy of Sciences, 117*, 19080–19091.
Szegedy, C, et al. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition* (pp. 2818–2826).
Tang, H, et al. (2018). Resting-state functional connectivity and deception: exploring individualized deceptive propensity by machine learning. *Neuroscience, 395*, 101–112.
Tang, H, et al. (2016). Interpersonal brain synchronization in the right temporo-parietal junction during face-to-face economic exchange. *Soc Cogn Affect Neurosci, 11*, 23–32.
Vaswani, A, et al. (2017). Attention is all you need. *Adv Neural Inf Process Syst, 30*.
Vecchione, M, et al. (2021). BIS and BAS sensitivities at different levels of personality description: A latent-variable approach with self-and informant-ratings. *J Pers Assess, 103*, 246–257.
Wang, H, et al. (2016). Functional brain network and multichannel analysis for the P300-based brain computer interface system of lying detection. *Expert Syst. Appl., 53*, 117–128.
Wang, Y, et al. (2024). 3D multi-modality Transformer-GAN for high-quality PET reconstruction. *Med Image Anal, 91*, Article 102983.
Wang, Y, et al. (2018). 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. *NeuroImage, 174*, 550–562.
Wang, Y, et al. (2019). 3D Auto-Context-Based Locality Adaptive Multi-Modality GANs for PET Synthesis. *IEEE Trans Med Imaging, 38*, 1328–1339.
Wang, Z, et al. (2022). Transformer model for functional near-infrared spectroscopy classification. *IEEE J Biomed Health Inform, 26*, 2559–2569.
Zhang, L, & Gläscher, J (2020). A brain network supporting social influences in human decision-making. *Sci. Adv., 6*, eabb4159.
Zhao, L, et al. (2023). Spatio-Temporal Variable Structure Graph Neural Network for EEG Data Classification. In *2023 6th International Symposium on Autonomous Systems (ISAS)* (pp. 1–6).