Decoding deceipt: EEG signatures of lying behavior under spontaneous versus instructed lying and truth-telling in a two-player game

Yiyu Chen
Korea University
Siamac Fazli
Nazarbayev University
Christian Wallraven (wallraven@korea.ac.kr)
Korea University

Research Article

Keywords: Electroencephalogram (EEG), single-trial decoding, event-related potential (ERP), deception, lie detection, event-related spectral perturbation (ERSP), decision-making

Posted Date: February 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2521275/v1

Additional Declarations: No competing interests reported.

RESEARCH

Decoding deceipt: EEG signatures of lying behavior under spontaneous versus instructed lying and truth-telling in a two-player game

Yiyu Chen YC[1], Siamac Fazli SF[2] and Christian Wallraven CW[1,3*]

*Correspondence: wallraven@korea.ac.kr
[1]Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea
[3]Department of Artificial Intelligence, Korea University, Seoul, South Korea
Full list of author information is available at the end of the article

Abstract

Background: Is it possible to decode lying behavior from neural signatures of the electroencephalography (EEG)? Existing studies on this topic have methodological limitations: tasks lack the incentive to lie, the act of lying is confounded with memory recollection, there is no sufficient distinction between instructed versus spontaneous decisions, and participants' risk-taking tendency is not controlled for. To address these limitations, we introduce a novel interactive, two-player game, where successful lying is incentivized in the reward scheme and that has both instructed and spontaneous conditions for participants matched by risk-taking tendency.

Methods: 24 participants were paired in the game according to their risk-taking tendency scores and measured using 32-channel EEG. Our multi-modal EEG analysis includes event-related potential (ERP), event-related spectral (ERS), and deep-learning-based single-trial decoding based on a one-dimensional convolutional neural network (1D-CNN).

Results: In ERP, two early components (P200 and N200) distinguished instructed truth-telling from other conditions, with a late component (N300) separating instructed lies from spontaneous conditions. Moreover, a late positive potential was indicative of spontaneous lying versus spontaneous truth-telling and was correlated with participants' risk-taking tendencies. In ERS, alpha and low beta were found to discriminate conditions. Importantly, we observed robust single-trial decoding performance under different conditions for the 1D-CNN. A gradient-based analysis further identified significant time-periods compatible with the ERP results from the classifier.

Conclusions: Our study represents the first effort to analyze EEG data not only through statistical properties but also with out-of-sample, deep-learning-based classification to decode deceit in the context of an iterative, game-theoretic experiment.

Keywords: Electroencephalogram (EEG); single-trial decoding; event-related potential (ERP); deception; lie detection; event-related spectral perturbation (ERSP); decision-making

Introduction

Lying is a ubiquitous component of human social interaction in which attempts to deceptively manipulate the recipient's belief are made, quite often for self-serving goals. Research on lie detection has drawn a substantial amount of research interest over the past decades with important applications within legal, moral, and clinical

domains. Reinforced by rapid developments in modern neuroimaging technology, such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), studies have made attempts to understand the underlying neural mechanisms of lying [1, 2, 3, 4, 5]

In contrast to early polygraph test approaches that focused on peripheral manifestations of lying, EEG and fMRI offer a more direct measure of neural processes related to lying and intentions to deceive. Over the past years, many such studies have investigated the feasibility of building lie detectors using, for example, the paradigm of a concealed information test (CIT, also known as Guilty Knowledge Test) [1]. The CIT facilitates psycho-physiological detection of prior knowledge, e.g. of specific details that would only be known to suspects involved in a potential crime. Therefore, a suspect would respond differently to pertinent items of that crime compared to irrelevant items, while innocent individuals would not reveal such differences between responses. Based on this protocol, EEG studies, for example, have found significant increases in the amplitude of the P3 signal triggered by pertinent stimuli among suspects (for a review, see [2]). Classification performed on P300 components found accuracy levels of up to 96.8% [6]. Source localization studies found a cluster of frontoparietal regions that are more engaged while lying compared to honest responses in CIT [3, 4, 5], which include regions such as the anterior cingulate cortex (ACC), the ventrolateral prefrontal cortex (VLPFC), and superior frontal gyrus (SFG). fMRI lie detection studies reported a weighted average ROC value of 0.94 [7] analyzing activity from these regions.

Although the accuracy of decoding lying behavior from neural activity seems impressive at first glance, the CIT-based lie detection design has limitations: firstly, as reported by several studies [8, 9], classification accuracy is generally reduced if participants are coached on countermeasures, such as focusing more on superficial aspects of the pertinent items, which helps to create "oddball" effects for irrelevant items. Secondly, the practical motivations to lie are absent in CIT compared to real-world situations, and it has been argued that instead of directly detecting lies (an impulsive and context-dependent behavior [10]), CIT rather detects differential responses based on the memory recognition of information [11]. As a consequence, the false-positive rate would be inevitably higher for innocent participants who were exposed to crime-related details [7, 12].

To mitigate these concerns, recent research has begun to investigate lying behavior using game-based designs. Lying is considered a type of decision-making behavior driven by extrinsic motivation [13] and involving cognitive control [14]. As such, it is often involved in games and usually reinforced by strategic countermeasures. In [15], the authors examined EEG neural correlates of spontaneous lying and instructed truthful decisions using a coin-guessing task with the former eliciting a higher amplitude of frontal N2 and smaller posterior P3 compared to the latter. In another study combining EEG and fMRI, spontaneous truth and spontaneous lies [16] were investigated. In their game, participants received varying amounts of monetary reward from an investor based on a proposal outlining how the money is distributed. Participants were asked to choose between the proposed honest option and a dishonest alternative to present the investor with while having a 50% risk of getting caught. Stronger fMRI activation was observed in the bilateral striatum and anterior insular for spontaneous lying choices. Similarly, fluctuations in ERP amplitudes

within a time window of $270 - 300$ milliseconds before the response were induced in the spontaneous lying condition. Such preparatory responses were also found in decreased motor-related components of ERPs for spontaneous lying in other tasks [17]. Conversely, other studies revealed no ERP differences for spontaneous truthful versus lying behaviors in interpersonal game tasks. In [18], an instructed truth control condition was added to the spontaneous conditions, and whereas both spontaneous conditions elicited larger fronto-central N300s and smaller P300s compared to the instructed condition, no difference was observed among spontaneous truthful and lying decisions. Finally, in [19] instructed and spontaneous truthful and lying responses were examined with only frontal N450 being linked to instructed truthful responses.

Overall, these heterogeneous results seem to suggest that spontaneous lying behavior may not mandate easily-detectable ERP components compared to spontaneous truthful behavior, lending support to the idea that spontaneous truthful actions can be used deceptively to mislead others [18, 19]. To date, only relatively few game-based lie detection EEG studies have been conducted, such that it is difficult to draw consistent conclusions. More specifically, one limitation of the aforementioned studies is that they have not tried to explicitly decode lying behavior itself. Concerning the experimental design, an additional point of concern is related to risk-taking: as is well-known, neural signatures of decision-making are highly dependent on a person's risk-taking tendency [20, 21], which in turn is often modulated by psychological arousal or stress [22]. Hence, proper control for the risk-taking behavior of participants is needed as they may directly affect the cognitive processes and behavioral dynamics of lying and truth-telling behavior. So far, such control is missing from existing studies.

In the present study, we, therefore, designed a two-person-based card game task that motivates real-world lying behavior. Specifically, we tested the full set of combinations of spontaneous/instructed and truthful/lying behaviors while measuring EEG. To emulate the dynamics of lying behavior in a more realistic manner, the two participants were assigned to a player and an observer role, where the objective of the player was to mislead the observer in order to win and gain higher rewards while the observer was incentivized to catch the player's potential lies. To control for the psychological arousal of all player-observer pairs, participants underwent the balloon analog risk-taking (BART) test [23] prior to the main experiment. Participants with similar scores, ages, and of the same gender were then paired up for the game. Finally, using the EEG data, we also performed single-trial classification to detect lies, focusing on the observer's point of view.

Method

Participants

24 participants (12 males and 12 females, aged 19-34, mean = 25 yrs, SD = $\pm$ 4.34) took part in the experiment. All had normal or correct-to-normal visual acuity and none of them had a history of neurological disease or injury. The participants were naïve to the card game paradigm and gave written informed consent before the start of the experiment and received payment of around 10US$ per hour for taking part in the study. The experiment was conducted in accordance with the

tenets of the Declaration of Helsinki and received IRB approval with the number KUIRB-2019-0043-01.

Apparatus

EEG was recorded with a total of 30 electrodes at a sampling frequency of 1000 Hz, using BrainAmp amplifiers and EasyCaps with an active electrode system (Brain Products, Munich, Germany). The measurements were performed with 30 EEG electrodes, namely: Fp2, F9,7,3,z,4,8,10, FC5,1,2,6, T7,8, C3,z,4, CP5,1,2,6, P7,3,z,4,8, PO3,4, O1,2. All EEG electrodes were nose-referenced and a forehead ground was used (Fpz). The impedance of all electrodes was kept below 10kΩ during the experiment. The setup time for the electrode configuration was 35 minutes on average.

All stimuli were presented on two 24″ monitors (LG, Seoul, South Korea) at a refresh rate of 60 Hz and a resolution of 1920px x 1080px. Participants' responses were collected using two RB-740 response pads (Cedrus Corporation, San Pedro, USA) with 6 buttons (number of $1-6$) used on one pad and 2 buttons ("Truth" and "Lie") used on the other pad. Participants' facial expressions were recorded using an HD pro C920 webcam (Logitech, Lausanne, Switzerland). The experiment was implemented in Python with PsychoPy [24]. Data preprocessing was performed with MATLAB (The MathWorks, Natick, MA, USA) using EEGLAB [25], and data analysis and classification were performed using the Berlin BCI toolbox [26]

Experimental Paradigm

In the experiment, participants were asked to play a deception card game with their counterparts. They were assigned the role of the player, while their counterpart was assigned the role of observer. Importantly, the player and observer were paired by similar risk-taking scores (measured by the BART paradigm), similar age, and same gender.

During the game, the player and observer sat opposite each other with two monitors in between them as shown in Figure 1b. At the start of each trial, players were assigned a card with a number on it and were then asked to tell the observers what number they saw on the card. Based on the player's facial expression and/or strategic considerations, observers then decided whether the information given to them by the player was a lie or the truth. In each trial, players had to adapt their behavior according to a response type cue, indicated by the card color. Color cues for each response type were randomly assigned to each participant. After the first session, the player and observer switched their roles to perform another session.

The game consisted of 11 rounds in total with 44 trials in each round. At the end of each round, there was a 30-second break before the next round started. Among the 44 trials, half were spontaneous trials, 11 trials instructed lies, and 11 trials instructed truths, with the order randomly shuffled. The stimuli in the card game consisted of cards with the numbers from $1-6$ in 3 different colors (black, purple, and blue) with the colors depending on the instruction.

The game started with explaining the card color cues to the player. As shown in Figure 1a, each trial then started with a 2-second fixation cross, followed by a card being displayed for 3 seconds at the center of the screen. The player was instructed to focus on the card and make a decision within 3 seconds, while their

facial expressions were shown to the observers in real time. Next, the player was asked to choose what card number to send to the observer with choices potentially limited by the color cue. Specifically, the player was to press a different number from the presented card in the "instructed lie" condition, press the same number in the "instructed truth" condition, and choose freely in the spontaneous condition. After the player pressed the corresponding button, a card in black with the player's chosen number was shown to the observer. The observer was then instructed to focus on the card for 3 seconds and to choose between "lie" or "truth". Following the observer's response, feedback with resulting scores or penalties for both player and observer was displayed on the screen. Scores were explained before the experiment and were designed to encourage lying behavior for the player and to catch lying behavior for the observer (success, in this case, was +15 points for the winner and -5 points for the loser); if the player was being honest, the winner received +10 points and the loser -5 points. The trial ended with a status screen showing the information on the total score earned, the number of won trials, won rounds, and the game progress.

EEG preprocessing

In the current study, only the player's data was analyzed. Data were downsampled to 100Hz, and a 0.1/49 Hz high-/low-pass filter was applied. Channel rejection was performed using EEGLab function *clean_artifact()* with channels whose line noise power was 4 standard deviations higher than their signals, and lower correlations than 0.85 with their reconstructed versions based on adjacent channels being rejected. With the same function, EEG data containing non-stationary, high-amplitude bursts were removed using artifact subspace reconstruction (ASR) [27], which is a principle-component-based method. The ASR procedure was applied using a 500-ms sliding window and a lax (20 standard deviations) threshold that removes extreme mechanical artifacts while preserving brain signal components. This method has been shown to improve the quality of a subsequent Independent Component Analysis (ICA) decomposition [28, 29]. Next, all removed channels were interpolated and EEG data were then re-referenced to a common average reference. ICA was performed and independent components (ICs) were subsequently separated into several signal categories (e.g., brain, muscle, eye, etc.) by a trained classifier ICLabel [30]. The ICs labeled as eye movements with probabilities higher than 0.7 were rejected.

Epochs of $3500ms$ were extracted starting at 500 ms before the stimulus presentation onset (the card with the colored number). Artifact-free epochs of each subject were grouped into four conditions, instructed truth (instT), instructed lie (instL), spontaneous truth (sponT), and spontaneous lie (sponL). The time interval between -500 to 0ms before the stimulus onset was used for the baseline correction of the epoch.

Event-related potential analysis

For the ERP analysis, the player's data was segmented from -500 ms to 3000 ms with respect to stimulus onset. Baseline correction was performed on the pre-stimulus interval from -500 ms to 0 ms. One player was excluded due to faulty EEG equipment and the remaining 23 participants were included for ERP analysis. Topographic maps of significant features were calculated by point-biserial correlation

coefficients [31], measuring the association of the trial type label to the electrode-wise ERP data. Using Fisher's transformation, correlations were transformed into unit variance z-scores for each subject, and grand average z-scores were obtained by weighted sums of individual z-scores over all subjects. In calculating grand-average statistics, inverse-variance weighting under a fixed-effects hierarchical model based on the sufficient statistics approach [32] was used. P-values for the hypothesis of zero correlation in the grand average were computed using a two-sided z-test. All reported P-values were Bonferroni-corrected to account for multiple hypothesis testing.

Event-related spectral analysis

For the time-frequency analysis, we computed ERSP as a measure of event-related dynamic changes in amplitude of the broad-band EEG frequency spectrum [33]. ERSP in the 4–50 Hz frequency range was computed in EEGLAB using Morlet wavelet decomposition starting with a 3-cycle wavelet at the lowest frequency, with a time resolution of 10 ms. ERSP values were common-baseline-corrected and converted to decibel units (dB) by log transformation and multiplying the log ratio with the factor 10. Therefore, the resulting ERSP shows the relative change in power in dB compared with the baseline during the event period. For the statistical testing, we used non-parametric Monte Carlo tests with 1000 randomization. Multiple comparison correction was done using cluster thresholding as implemented in Fieldtrip [34] (clusteralpha: 0.05; maxsum as the dependent variable of the clustering).

One-dimensional convolution neural network classification

To test single-trial decoding, we used a 10-layer one-dimensional convolution neural network (1D-CNN) [35] in a stratified ten-fold cross-validation scheme. Such a one-dimensional CNN was found helpful for extracting important local features between adjacent element values of the feature vector[36]. Our particular network has been shown to outperform previous state-of-the-art CNN methods on motor imagery data; performance has been boosted through an electrode selection approach using pairs of symmetrical electrodes located at the region of interest [35, 37]. In the present work, we selected electrode pairs based on the significant scalp pattern of ERP results. As shown in figure 2, a total of 17 pairs of electrodes were used for the network training, these electrode pairs include frontal-occipital and X-pattern symmetrical electrodes with respect to transverse line across electrode T7-C3-Cz-C4-T8.

To test whether the classification results of the 1D-CNN model were statistically significant above the chance level, exact binomial tests were conducted at within-participant level and for each of the six experimental condition contrasts. The proportion of correct and false predictions was compared to a null model with a prediction accuracy of 0.5 (chance level).

To gain a better understanding of the trained 1D-CNN classifier and explore its learned feature pattern on the time dimension, a gradient class activation map (Grad-CAM) approach was adopted [38] to provide interpretable values to assess the impact of different time points. The calculated feature importances using Grad-CAM were normalized for each training in ten-fold cross-validation.

Results

In the following, we use the abbreviations of sponL = spontaneous lie, sponT = spontaneous truth, instL = instructed lie, and instT = instructed truth for the four types of possible decisions of the observer.

Behavioral Results

In the spontaneous conditions, participants were able to decide whether they would tell the truth or they would lie. In our sample, we found that participants (N=23) made a significantly higher number of truthful decisions ($mean = 132, SD = \pm 20$) compared to lie decisions ($mean = 109, SD = \pm 20$; $t(22) = -3.809$, $p < 0.001$). No significant correlation was found between the individual percentage of lies in the spontaneous condition and their BART scores ($r = -0.32$, $p = 0.13$).

For the reaction times, previous studies have shown that the reaction data resembles a convolution of normal and exponential distributions (Ex-Gaussian)[39]. Data transformation using Box-Cox has been shown to reduce the deviation from normality to meet the normality assumption for parametric statistic tests [39]. Our reaction time data were therefore Box-Cox transformed with $\lambda = 0.3$ for the following statistical analysis. A 4-level (condition type: instL, instT, sponL, sponT) one-way repeated measure ANOVA revealed a significant main effect of condition type, $F(3, 66) = 4.8, p = 0.0044, \eta_p^2 = 0.18$ (see Figure 2 in supplementary material). Through paired t-tests as a follow-up analysis, we found that participants responded significantly slower in the sponL condition ($GM_{untransfromed} = 552ms$) compared to all other conditions (instT: $GM_{untransfromed} = 513ms$, $t_{(sponL,instT)}(22) = 2.95$, $p = 0.0073$; instL: $GM_{untransfromed} = 521ms$, $t_{(sponL,instL)}(22) = 2.97$, $p = 0.0070$; sponT: $mean = 502ms$, $t_{(sponL,sponT)}(22) = 3.35$, $p = 0.0029$). No significant differences were observed for other condition pairs (see Figure 3a,b).

Event-related potential analysis

Next, we investigated whether the stimulus-locked event-related potentials (ERPs) from the four different conditions were statistically different from one another. We observed significant differences between conditions in multiple components spanning different intervals within the ERP timeframe: P200, N200, N300, late positive potential (LPP), and post-LPP intervals (see Figure 4).

When contrasting to the instT condition, the other three conditions elicit a more pronounced P200 response during 170-200 ms and N200 during 240-290ms in figure 4a. P200 was observed to be more fronto-central located while N200 was more prefrontal focused in their scalp map (figure 4b). Another strong negative component was found between 320 and 340 ms (N300) for instL compared to instT and spontaneous conditions (figure 4a). This N300 was observed in the central area of the scalp. Additionally, a prolonged prefrontal-centered LPP from 490 to 520 ms was observed when contrasting instT with the other three conditions, with a more pronounced prefrontal positivity at the post-stage of LPP (post-LPP) further distinguishing these three conditions. This prefrontal post-LPP was highest for sponL followed by sponT and instL. For the mixed comparison of instL vs. sponT and instT vs. sponL, please refer to the supplementary materials.

Event-related spectral analysis

Next, we examined the different conditions in the time-frequency domain through event-related spectral perturbation (ERSP). Generally, more significant clusters were found in alpha and low beta at the posterior electrodes between 500ms to 900ms (see Figure 5). The contrast between the two truth-telling conditions (sponT vs. instT) also showed significant alpha and low beta in the fronto-central region. This activation, however, was absent in the spontaneous contrast (sponL vs. sponT), which showed only significant gamma contrast changes at later time periods in the central-parietal region.

BART Risk-taking tendency

To investigate potential relationships between the participant's risk-taking tendency and the neural activity in the decision-making process during the game, we performed a correlation analysis between participants' BART scores and the component amplitude differences between sponL and sponT (as those contrast the voluntary risk-taking decisions). As indicated in Figure 4, significant component-wise differences between sponT and sponL were observed only in post-LPP intervals, which is why we used this specific component for the correlation analysis. Specifically, a cluster of electrodes marked as significant in the post-LPP interval was chosen for the analysis, including prefrontal, frontal, and fronto-central electrodes. For this cluster, we observed a significant negative correlation at $r = -0.55$, $p = 0.0066$ (see Figure 4c), indicating that participants with a higher risk-taking tendency tend to show a reduced frontal post-LPP ERP.

One-dimensional convolution neural network classification

The mean cross-validated classification accuracy resulting from One-dimensional convolution neural network (1D-CNN) for all binary combinations of conditions is shown in Figure 6a, T-tests between subject-wise accuracy and chance level determined that binary classification performance was above chance for instL vs. instT ($mean = 0.54$, $SD = \pm 0.09$, $t(22) = 2.87$, $p < 0.001$), sponT vs. instT ($mean = 0.55$, $SD = \pm 0.07$, $t(22) = 4.39$, $p < 0.0001$), sponL vs. instT ($mean = 0.54$, $SD = \pm 0.07$, $t(22) = 4.49$, $p < 0.0001$), sponT vs. instL ($mean = 0.54$, $SD = \pm 0.08$, $t(22) = 3.48$, $p < 0.005$), sponL vs. instL ($mean = 0.54$, $SD = \pm 0.09$, $t(22) = 3.09$, $p < 0.005$), sponL vs. sponT ($mean = 0.54$, $SD = \pm 0.07$, $t(22) = 4.45$, $p < 0.0001$).

To identify discriminative time windows for decoding the binary combinations of conditions, we performed Grad-CAM on the trained 1D-CNN classifier. As shown in figure 6b, the main important features are located at 0-300ms followed by another later peak interval between 1000-1500ms, suggesting classification performance was strongly influenced by these two components.

Discussion

The present study employed a two-person card game and investigated the feasibility of decoding the intent of truthful or lying behavior of the active player under instructed or spontaneous circumstances based on EEG data. Our behavioral data showed that players in the spontaneous condition made around 30% more truthful than lying decisions. At the same time, the response time for the spontaneous

lie trials was slower compared to the spontaneous truth trials. This result can be understood from previous research showing that choices requiring more cognitive activity will result in longer response times compared to choices that involve an instinctive response [40]. In the analysis of ERP, we observed significant frontal and parietal differences, which will be discussed in more detail in the following. Furthermore, alpha and low beta were found to discriminate conditions in ERS. Finally, the decoding of several conditions at the single-trial level was possible above chance, predicting subsequent behavior. In the following, we will first put the ERP results into context, followed by a brief discussion of the ERS and decoding results.

Event-related potentials

Concerning the results for ERP components, separating instT from other conditions, we observed significant P200 differences in fronto-central areas - this distinction was not present for other pairs. The P200 observed in the present study was accompanied by bilateral posterior negativity similar to a proposed network for control of visual attention [41, 42]: in this model, the ACC included medial frontal regions, which are part of a voluntary attention system that determines where and how attention should be directed to meet the demands of a particular task, whereas the posterior parietal regions then engage attention. Recent studies confirmed enhanced fronto-central P200s to be related to emotional salience [43], risky information [44], or mismatch [45]. In our experiment, participants' responses in the instructed truth condition are limited to a single button press, which requires very little attention, whereas instructed lie and both spontaneous conditions require participants to choose a response button to perform lying or truth-telling, hence providing a potential explanation for the observed differences.

Following the early processing component of the P200, a frontal N200 is elicited at higher levels for the instT condition. This ERP component is primarily found in go/no-go studies and is theorized to index cognitive control involved in attentional conflict monitoring [46, 47] (for a review see [48]). The increased amplitude of an N200 is related to trials requiring higher demand for cognitive control or response monitoring [49, 50]. Previous studies found the major neural generator of the N200 to be in the left anterior region of the midcingulate cortex and inferior frontal region, including in the executive attention network [51, 49]. Moreover, since performing a deceptive response requires more cognitive control [52], this N200 conflict monitoring effect has been linked to deception intentions in various tasks of interpersonal deception where the opponent had to guess the correctness of a player's response [18, 19, 53]: increased frontal-central activity around 200ms was elicited in both instructed deception and spontaneous conditions compared to an instructed truth condition in [19] and in [18] when comparing a spontaneous condition to the instructed truth condition. Our results are in agreement with these previous studies, showing a consistently more negative N200 in both an instructed lie and the two spontaneous conditions relative to the instructed truth condition, a finding that is potentially explained by higher levels of cognitive control required for these former conditions.

In addition to the strong frontal N200 response contrasting instructed truth to instructed lie and the spontaneous conditions, a subsequent, more centrally-located

N300 difference was observed contrasting only the instructed lie versus the two spontaneous conditions. Several studies have found that a concurrent N200-N300 effect is linked to different levels of semantic processing with the N200 being related to more perceptual processes whereas the accompanying N300 would rather be associated with semantic integration processing. For example, an increased N300 effect was observed for inconsistent objects in inverted scenes, whereas the N200 disambiguated these only for upright scenes [54]. Similarly, an N300 effect was observed for a subordinated within-condition, whereas the N200 was only found for subordinated between-conditions[55]. Hence, during stimulus identification, the N200 does not differentiate fully between semantically-related pairs [56, 57], and the N300 is sensitive to more specific information in the presence of semantic incongruencies. In our case, the N200 may therefore differentiate between the cue-related default behavior (instructed truth) and other behaviors, whereas the N300 shows a more fine-grained distinction between cue-related processing of forced-choice deception (instructed lie) versus self-determined deception (spontaneous) with an increased amplitude for the forced-choice behavior. Similar to [58], we also observe a mid-central difference in a negative component between 300 to 400ms.

For the instructed truth condition, we observed a sustained, more positive frontal LPP accompanied by posterior negative deflections for instL, sponT, and sponL compared to instT. A number of previous studies have found the LPP to be related to decisions involving ambiguous choices in perceptual decision-making tasks [59, 60], and in relation to decision uncertainty in memory retrieval [61]. Source estimations suggested that this LPP arises from a distributed brain network, including ACC, posterior cingulate cortex (PCC), and insula. This network is characterized by electro-cortical stimulation of the prefrontal cortex during attentional arousal modulation of the LPP over the central parietal electrode sites [62, 59]. In our experiment, lying is characterized by choosing one out of five untrue responses in order to mislead the opponent, spontaneous truth-telling involves the prior resolution of choosing between a lie and a truth, whereas instructed truth-telling merely involves a single button press with the corresponding number given by the cue. Hence, the ambiguity of the choices is graded from higher to lower in our conditions, which fits with the suggested interpretation of the LPP as a decision ambiguity index.

Furthermore, a subsequent prefrontal, higher post-LPP was observed in the two spontaneous conditions compared to instL. It occurs at a later stage of LPP, and this delay could be due to the increased amount of time required for instructed lying and spontaneous decisions compared to instructed truth. As mentioned earlier, a decision for the instructed lie condition requires participants to choose one out of five untrue numbers, whereas the spontaneous conditions require the first choice between truth and lie, before proceeding to a specific number choice, indicating higher decision ambiguity for the spontaneous conditions. Here, the frontal effect size in the sponT condition is small (two electrodes in the left frontal area), which could be due to the decision delay - effect sizes in the posterior regions seem larger, however, which may indicate that the decision ambiguity for sponT could be slightly stronger than that for instL.

Finally, a more pronounced post-LPP component was observed in prefrontal regions for sponL relative to sponT. Similar to the previous discussion for this component in the instL and instT conditions, a higher decision ambiguity may be the reason here. However, as early components of the LPP are more perceptually driven, and the actual cue for the sponL and sponT is not visually different, we do not observe any significant P200, N200, and N300 components that differentiate sponL and sponT. This result is consistent with previous studies [18, 19, 17].

We also observed a negative correlation between the frontal post-LPP voltage contrasting sponL and sponT and the BART risk-taking score. Hence, individuals with a lower risk-taking tendency may feel a greater decision ambiguity between lying and telling the truth, as choosing to lie in our game is related to a higher penalty if caught.

Event-related spectral analysis

In the time-frequency domain, we observed large-scale, significant differences in the alpha and low beta range for all contrasts except for sponL vs. sponT. These alpha and low beta differences happened mainly in posterior regions and frontal regions in sponT vs. instT. The corresponding spectral range has previously been associated with attentional modulation [63, 64], confirming the distinct attentional allocation differences discussed above for the early time periods in our task. Similarly, the absence of such differences in the spontaneous contrast (sponL vs. sponT) fits with our general observation that timing associated with decision-making in these two conditions is critically different.

Single-trial classification

In the decoding analysis, we investigated single-trial classification using 1D-CNN to functionally relate EEG features to behavioral performance and obtained above-chance classification accuracy for all six binary classification pairs. Importantly, the Grad-CAM shows early time points containing the main discriminative features for 1D-CNN to make classification decisions. This confirms that early ERP components (P200, N200, and N300) are able to classify the truths and lies under both instructed and spontaneous conditions via the 1D-CNN.

Ethics approval and consent to participate
All participants gave written informed consent before the start of the experiment. The experiment was conducted in accordance with the tenets of the Declaration of Helsinki and received IRB approval with the number KUIRB-2019-0043-01.

Consent for publication
Not applicable.

Availability of data and materials
The code and dataset analyzed during the current study are available from the corresponding author on reasonable request due to the need for a formal data-sharing agreement.

Competing interests
The authors declare that they have no competing interests.

Funding
This study was supported by the National Research Foundation of Korea under project BK21 FOUR and grants NRF-2017M3C7A1041824, NRF-2019R1A2C2007612, the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (No. 2017-0-00451, Development of BCI based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning; No. 2019-0-00079, Department of Artificial Intelligence, Korea University; No. 2021-0-02068, Artificial Intelligence Innovation Hub), and the Nazarbayev University Faculty-Development Competitive Research Grants Program (240919FD3926).

Authors' contributions

All authors conceptualized the experimental design and analysis, and wrote and reviewed the manuscript. SF provided initial funding and supervision. CW provided funding, supervision, and infrastructure. YC contributed the main parts of the coding and analysis.

Author details

[1]Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea. [2]Department of Computer Science, Nazarbayev University, Nur-Sultan, Kazakhstan. [3]Department of Artificial Intelligence, Korea University, Seoul, South Korea.

References

1. Lykken, D.T.: The GSR in the detection of guilt. Journal of Applied Psychology **43**(6), 385 (1959)
2. Rosenfeld, J.P.: P300 in detecting concealed information. Memory detection: Theory and application of the Concealed Information Test, 63–89 (2011)
3. Christ, S.E., Van Essen, D.C., Watson, J.M., Brubaker, L.E., McDermott, K.B.: The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta-analyses. Cerebral Cortex **19**(7), 1557–1566 (2009). doi:10.1093/cercor/bhn189
4. Farah, M.J., Hutchinson, J.B., Phelps, E.A., Wagner, A.D.: Functional MRI-based lie detection: Scientific and societal challenges. Nature Reviews Neuroscience **15**(2), 123–131 (2014). doi:10.1038/nrn3665
5. Lisofsky, N., Kazzer, P., Heekeren, H.R., Prehn, K.: Investigating socio-cognitive processes in deception: A quantitative meta-analysis of neuroimaging studies. Neuropsychologia **61**, 113–122 (2014). doi:10.1016/j.neuropsychologia.2014.06.001
6. Bablani, A., Edla, D.R., Tripathi, D., Dodia, S., Chintala, S.: A synergistic concealed information test with novel approach for EEG channel selection and SVM parameter optimization. IEEE Transactions on Information Forensics and Security **14**(11), 3057–3068 (2019)
7. Peth, J., Sommer, T., Hebart, M.N., Vossel, G., Büchel, C., Gamer, M.: Memory detection using fMRI—does the encoding context matter? NeuroImage **113**, 164–174 (2015)
8. Hsu, C., Begliomini, C., Dall'Acqua, T., Ganis, G.: The effect of mental countermeasures on neuroimaging-based concealed information tests. Human brain mapping **40**(10), 2899–2916 (2019)
9. Lukács, G., Weiss, B., Dalos, V.D., Kilencz, T., Tudja, S., Csifcsák, G.: The first independent study on the complex trial protocol version of the p300-based concealed information test: Corroboration of previous findings and highlights on vulnerabilities. International Journal of Psychophysiology **110**, 56–65 (2016)
10. Ganis, G., Keenan, J.P.: The cognitive neuroscience of deception. Social Neuroscience **4**(6), 465–472 (2009)
11. Kleinberg, B., Verschuere, B.: Memory detection 2.0: The first web-based memory detection test. PloS one **10**(4), 0118715 (2015)
12. Winograd, M.R., Rosenfeld, J.P.: The impact of prior knowledge from participant instructions in a mock crime P300 concealed information test. International journal of psychophysiology **94**(3), 473–481 (2014)
13. Linke, J., Kirsch, P., King, A.V., Gass, A., Hennerici, M.G., Bongers, A., Wessa, M.: Motivational orientation modulates the neural response to reward. Neuroimage **49**(3), 2618–2625 (2010)
14. Greene, J.D., Paxton, J.M.: Patterns of neural activity associated with honest and dishonest moral decisions. Proceedings of the National Academy of Sciences **106**(30), 12506–12511 (2009)
15. Hu, X., Pornpattananangkul, N., Nusslock, R.: Executive control-and reward-related neural processes associated with the opportunity to engage in voluntary dishonest moral decision making. Cognitive, Affective, & Behavioral Neuroscience **15**(2), 475–491 (2015)
16. Sun, D., Lee, T.M., Wang, Z., Chan, C.C.: Unfolding the spatial and temporal neural processing of making dishonest choices. PloS one **11**(4), 0153660 (2016)
17. Panasiti, M.S., Pavone, E.F., Mancini, A., Merla, A., Grisoni, L., Aglioti, S.M.: The motor cost of telling lies: Electrocortical signatures and personality foundations of spontaneous deception. Social neuroscience **9**(6), 573–589 (2014)
18. Sai, L., Wu, H., Hu, X., Fu, G.: Telling a truth to deceive: Examining executive control and reward-related processes underlying interpersonal deception. Brain and cognition **125**, 149–156 (2018)
19. Carrión, R.E., Keenan, J.P., Sebanz, N.: A truth that's told with bad intent: An erp study of deception. Cognition **114**(1), 105–110 (2010)
20. Sacré, P., Kerr, M.S., Subramanian, S., Fitzgerald, Z., Kahn, K., Johnson, M.A., Niebur, E., Eden, U.T., González-Martínez, J.A., Gale, J.T., *et al.*: Risk-taking bias in human decision-making is encoded via a right–left brain push–pull system. Proceedings of the National Academy of Sciences **116**(4), 1404–1413 (2019)
21. Chen, Y., Wallraven, C.: Pop or not? EEG correlates of risk-taking behavior in the balloon analogue risk task. In: 2017 5th International Winter Conference on Brain-Computer Interface (BCI), pp. 16–19 (2017). IEEE
22. Jordan, J., Sivanathan, N., Galinsky, A.D.: Something to lose and nothing to gain: The role of stress in the interactive effect of power and stability on risk taking. Administrative Science Quarterly **56**(4), 530–558 (2011)
23. Lejuez, C.W., Read, J.P., Kahler, C.W., Richards, J.B., Ramsey, S.E., Stuart, G.L., Strong, D.R., Brown, R.A.: Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (BART). Journal of Experimental Psychology: Applied **8**(2), 75 (2002)
24. Peirce, J.W.: Psychopy—psychophysics software in python. Journal of neuroscience methods **162**(1-2), 8–13 (2007)
25. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods **134**(1), 9–21 (2004)
26. Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., Maeder, C., Ramsey, L.E., Sturm, I., Curio, G., *et al.*: The berlin brain–computer interface: non-medical uses of BCI technology. Frontiers in neuroscience **4**, 198 (2010)
27. Kothe, C.A.E., Jung, T.-p.: Artifact removal techniques with signal reconstruction. Google Patents. US Patent App. 14/895,440 (2016)

28. Chang, C.-Y., Hsu, S.-H., Pion-Tonachini, L., Jung, T.-P.: Evaluation of artifact subspace reconstruction for automatic artifact components removal in multi-channel eeg recordings. IEEE Transactions on Biomedical Engineering **67**(4), 1114–1121 (2019)
29. Artoni, F., Fanciullacci, C., Bertolucci, F., Panarese, A., Makeig, S., Micera, S., Chisari, C.: Unidirectional brain to muscle connectivity reveals motor cortex control of leg muscles during stereotyped walking. Neuroimage **159**, 403–416 (2017)
30. Pion-Tonachini, L., Kreutz-Delgado, K., Makeig, S.: Iclabel: An automated electroencephalographic independent component classifier, dataset, and website. NeuroImage **198**, 181–197 (2019)
31. Tate, R.F.: Correlation between a discrete and a continuous variable. point-biserial correlation. The Annals of mathematical statistics **25**(3), 603–607 (1954)
32. Dowding, I., Haufe, S.: Powerful statistical inference for nested data using sufficient summary statistics. Frontiers in human neuroscience **12**, 103 (2018)
33. Makeig, S.: Auditory event-related dynamics of the eeg spectrum and effects of exposure to tones. Electroencephalography and clinical neurophysiology **86**(4), 283–293 (1993)
34. Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M.: Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. Computational intelligence and neuroscience **2011** (2011)
35. Mattioli, F., Porcaro, C., Baldassarre, G.: A 1D CNN for high accuracy classification and transfer learning in motor imagery eeg-based brain-computer interface. Journal of Neural Engineering **18**(6), 066053 (2022)
36. Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for EEG decoding and visualization. Human brain mapping **38**(11), 5391–5420 (2017)
37. Lun, X., Yu, Z., Chen, T., Wang, F., Hou, Y.: A simplified cnn classification method for mi-eeg via the electrode pairs signals. Frontiers in Human Neuroscience **14**, 338 (2020)
38. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
39. Marmolejo-Ramos, F., Cousineau, D., Benites, L., Maehara, R.: On the efficacy of procedures to normalize ex-gaussian distributions. Frontiers in psychology **5**, 1548 (2015)
40. Rubinstein, A.: Instinctive and cognitive reasoning: A study of response times. The Economic Journal **117**(523), 1243–1259 (2007)
41. Posner, M.I., Dehaene, S.: Attentional networks. Trends in neurosciences **17**(2), 75–79 (1994)
42. Bigman, Z., Pratt, H.: Time course and nature of stimulus evaluation in category induction as revealed by visual event-related potentials. Biological psychology **66**(2), 99–128 (2004)
43. Chou, L.-C., Pan, Y.-L., Lee, C.-l.: Emotion anticipation induces emotion effects in neutral words during sentence reading: Evidence from event-related potentials. Cognitive, Affective, & Behavioral Neuroscience **20**(6), 1294–1308 (2020)
44. Ma, Q., Jin, J., Wang, L.: The neural process of hazard perception and evaluation for warning signal words: evidence from event-related potentials. Neuroscience letters **483**(3), 206–210 (2010)
45. Xiao, F., Sun, T., Qi, S., Chen, Q.: Common and distinct brain responses to detecting top-down and bottom-up conflicts underlying numerical inductive reasoning. Psychophysiology **56**(12), 13455 (2019)
46. Falkenstein, M., Hoormann, J., Hohnsbein, J.: ERP components in Go/Nogo tasks and their relation to inhibition. Acta psychologica **101**(2-3), 267–291 (1999)
47. Nieuwenhuis, S., Yeung, N., Van Den Wildenberg, W., Ridderinkhof, K.R.: Electrophysiological correlates of anterior cingulate function in a Go/No-go task: effects of response conflict and trial type frequency. Cognitive, affective, & behavioral neuroscience **3**(1), 17–26 (2003)
48. Folstein, J.R., Van Petten, C.: Influence of cognitive control and mismatch on the N2 component of the ERP: a review. Psychophysiology **45**(1), 152–170 (2008)
49. Enriquez-Geppert, S., Konrad, C., Pantev, C., Huster, R.J.: Conflict and inhibition differentially affect the N200/P300 complex in a combined go/no-go and stop-signal task. Neuroimage **51**(2), 877–887 (2010)
50. Hu, X., Pornpattananangkul, N., Rosenfeld, J.P.: N200 and P300 as orthogonal and integrable indicators of distinct awareness and recognition processes in memory detection. Psychophysiology **50**(5), 454–464 (2013)
51. Huster, R., Westerhausen, R., Pantev, C., Konrad, C.: The role of the cingulate cortex as neural generator of the N200 and P300 in a tactile response inhibition task. Human brain mapping **31**(8), 1260–1271 (2010)
52. Nunez, J.M., Casey, B., Egner, T., Hare, T., Hirsch, J.: Intentional false responding shares neural substrates with response conflict and cognitive control. Neuroimage **25**(1), 267–277 (2005)
53. Zeki, S., Goodenough, O., Spence, S.A., Hunter, M.D., Farrow, T.F., Green, R.D., Leung, D.H., Hughes, C.J., Ganesan, V.: A cognitive neurobiological account of deception: evidence from functional neuroimaging. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **359**(1451), 1755–1762 (2004)
54. Lauer, T., Willenbockel, V., Maffongelli, L., Võ, M.L.-H.: The influence of scene and object orientation on the scene consistency effect. Behavioural Brain Research **394**, 112812 (2020)
55. Hamm, J.P., Johnson, B.W., Kirk, I.J.: Comparison of the N300 and N400 ERPs to picture stimuli in congruent and incongruent contexts. Clinical Neurophysiology **113**(8), 1339–1350 (2002)
56. Sitnikova, T., Holcomb, P.J., Kiyonaga, K.A., Kuperberg, G.R.: Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. Journal of cognitive neuroscience **20**(11), 2037–2057 (2008)
57. Yum, Y.N., Holcomb, P.J., Grainger, J.: Words and pictures: An electrophysiological investigation of domain specific processing in native chinese and english speakers. Neuropsychologia **49**(7), 1910–1922 (2011)
58. Wu, H., Hu, X., Fu, G.: Does willingness affect the N2-P3 effect of deceptive and honest responses? Neuroscience letters **467**(2), 63–66 (2009)
59. Sun, S., Zhen, S., Fu, Z., Wu, D.-A., Shimojo, S., Adolphs, R., Yu, R., Wang, S.: Decision ambiguity is mediated by a late positive potential originating from cingulate cortex. NeuroImage **157**, 400–414 (2017)

60. Sun, S., Yu, R., Wang, S.: A neural signature encoding decisions under perceptual ambiguity. eneuro **4**(6) (2017)
61. Graziano, M., Parra, L.C., Sigman, M.: Neural correlates of perceived confidence in a partial report paradigm. Journal of Cognitive Neuroscience **27**(6), 1090–1103 (2015)
62. Yoder, K.J., Decety, J.: Spatiotemporal neural dynamics of moral judgment: a high-density ERP study. Neuropsychologia **60**, 39–45 (2014)
63. Klimesch, W.: Alpha-band oscillations, attention, and controlled access to stored information. Trends in cognitive sciences **16**(12), 606–617 (2012)
64. Wróbel, A., Ghazaryan, A., Bekisz, M., Bogdan, W., Kamiński, J.: Two streams of attention-dependent β activity in the striate recipient zone of cat's lateral posterior–pulvinar complex. Journal of Neuroscience **27**(9), 2230–2240 (2007)

Figures

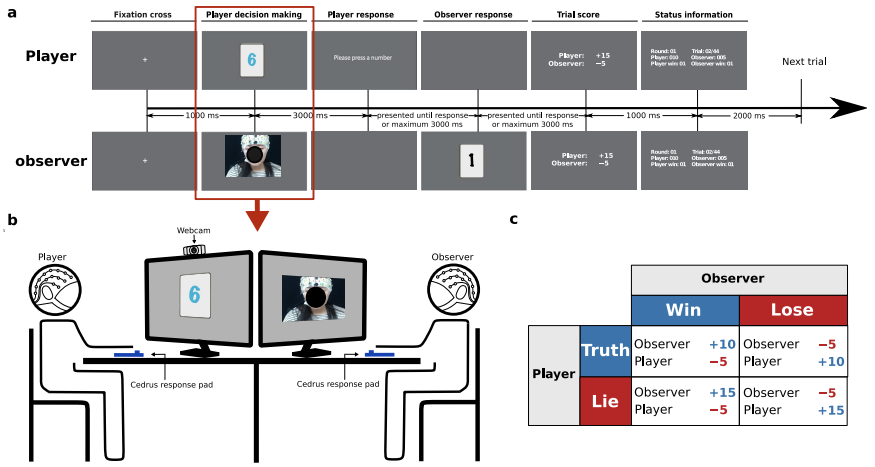

Figure 1 Experimental setup. a Trial structure of the experiment. The trial started with a 1-second fixation cross (1st figure), followed by the player's 3 seconds decision-making phase where their face was shown to the observer while looking at the card (2nd figure). Next, the player was instructed to give the response with a maximum reaction time of 3 seconds (3rd figure). The player's response was then shown to the observer who was asked to decide between lie or truth within a maximum response period of 3 seconds (4th figure). The trial ended with a 1-second trial score and 2 seconds status information screen (5th and 6th figure). **b** Experimental setup with a schematic illustration of the situation in the player's decision-making phase. The player and observer sat face to face with monitors in between them. Participants' responses were collected using a response pad and the player's face was displayed in real-time to the observer via a webcam stream. **c** Payoff matrix of the game for player and observer.

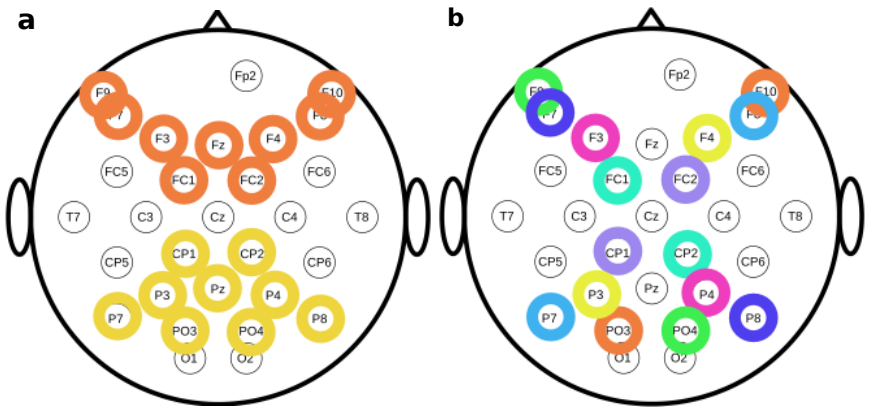

Figure 2 Selection of electrode pairs. a) fronto-occipital symmetrical electrode pairs **b)** X-pattern electrode pairs.

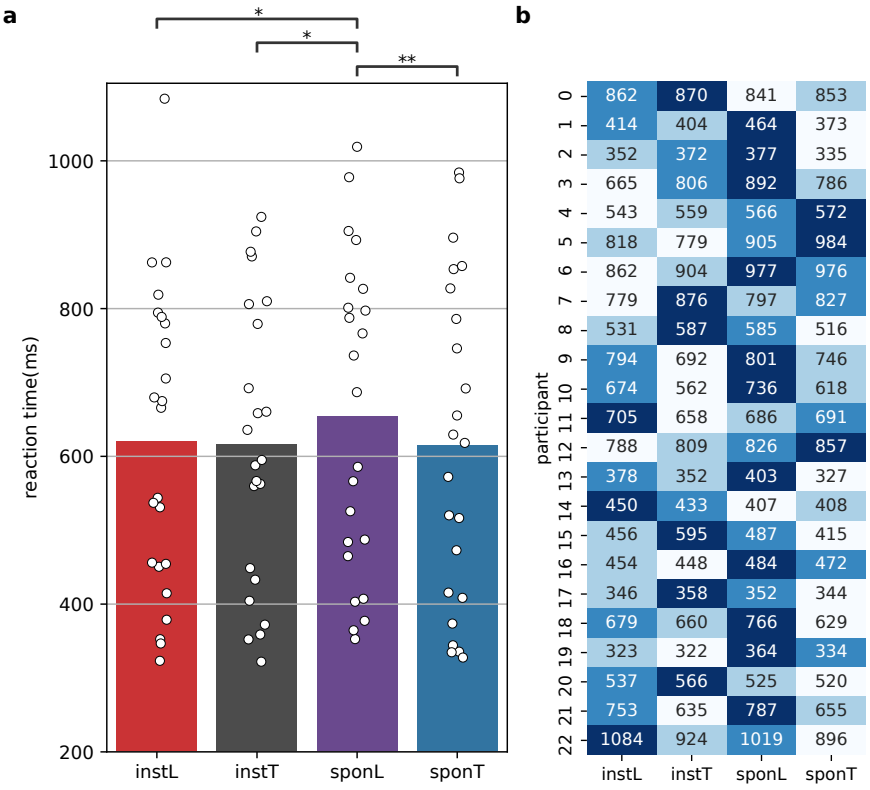

Figure 3 Reaction time. a) Barplot of reaction times for all 4 conditions across all participants. **b**) Heatmap of individual reaction times in milliseconds with the x-axis showing the condition and the y-axis showing the participant; deeper color indicates slower reaction time compared to other conditions for each participant.

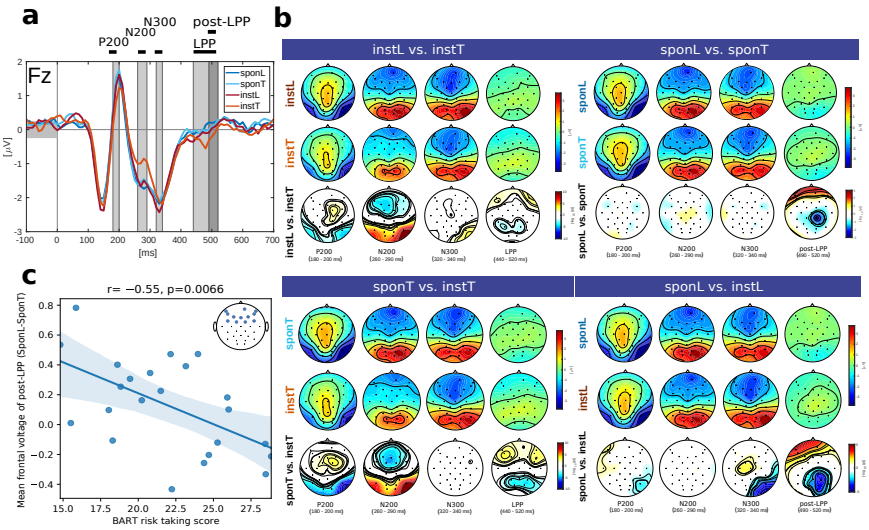

Figure 4 Event-related potential. a) grand average ERP analysis of all conditions for electrodes Fz. Shaded areas indicate P200, N200, N300, LPP, and post-LPP (overlapped with LPP), respectively; **b)** shows the ERP scalp map (rows 1 and 2) and signed logarithm p-values (row 3) indicating the grand average statistical significance of difference for instL vs. instT (left top), sponL vs. instT (right-top), sponT vs. instT (left-bottom), and sponL vs. sponT (right bottom). Bold contours indicate $p < 0.05$, where p-values are corrected for multiple comparisons. **c)** Correlation between the BART risk-taking score and mean frontal voltage. The mean frontal voltage was calculated using the data in the post-LPP component contrasting sponL and sponT. Translucent bands around the regression line indicate the 95% confidence interval for the regression estimate. Electrodes used for correlation are shown top right.

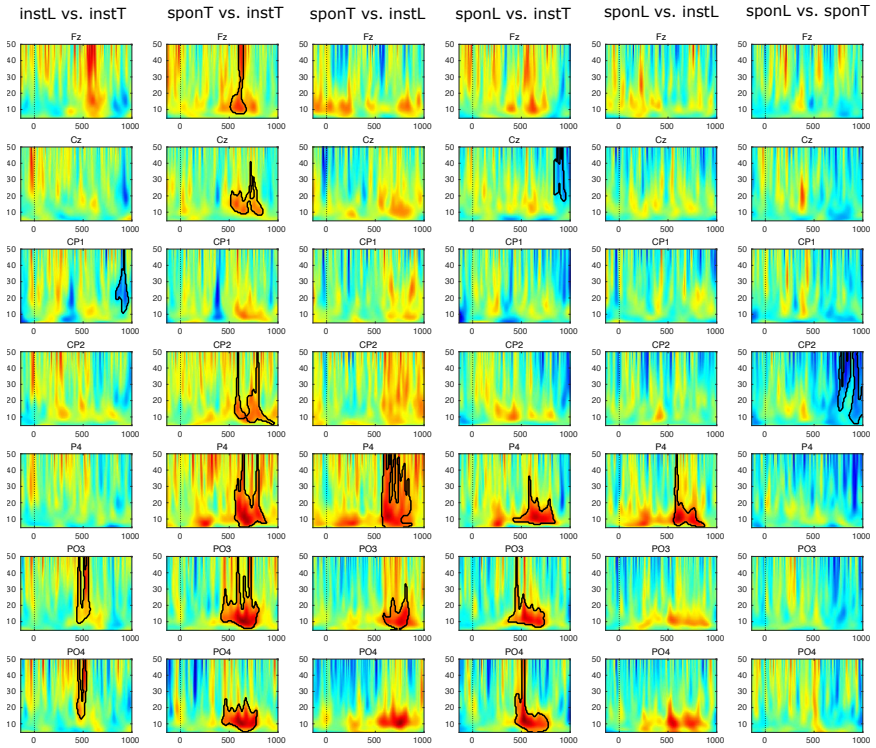

Figure 5 Time-frequency results. Columns index the six binary combinations of condition contrasts, and row indexes selected channels. Within each time-frequency map, the x-axis indicates the time in ms, and the y-axis indicates the frequency in Hz; significant differences for each contrast are contoured by black lines at $p < 0.05$.

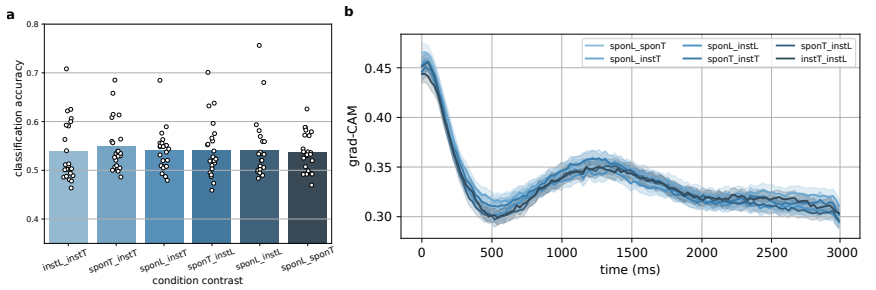

Figure 6 1D-CNN single-trial decoding. a) The mean cross-validated 1D-CNN classification accuracy for all six binary combinations of conditions using 0 to 3000 ms post-stimulus data. Individual accuracy is indicated by white dots. **b)** Feature importance calculated using Grad-CAM for all six binary combinations of condition

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalforChenetal.pdf