Decoding Participant State During False Belief Task

David Cremins[1], Sophia Sun[1,2], Corinne Donnay[3], and Shannon Klotz[4]

1. Cognitive Science and 2. Linguistics, Pomona College; 3. Computational Neuroscience and 4. Cognitive Neuroscience, Scripps College

Introduction

Theory of Mind (ToM) is the ability to attribute mental states to others and oneself, and is associated with passive judgments about another's belief and perceptions. Previous research shows two systems of mental computation underlying ToM: explicit ToM (deliberately considering another's mental state) and implicit ToM (considering another's mental state without engaging in deliberate reflection). These form the *two-system account of ToM*[1]. However, the extent of their integration still remains unknown. Common methods of studying ToM use variations of classic location-change false belief tasks (FBT) (e.g., Sally-Ann). This task measures whether participants can understand that other individuals might have a false belief that corresponds to the way the world actually is. While many false belief tasks do not sufficiency acknowledge the role of linguistic capabilities, perspective taking, and social understanding, variations of this task show implicit ToM at different levels of development. The *Interactive False Belief Teacher-Learner Task* used in this study incorporates both explicit (prediction) and implicit (choice) ToM. The task was developed to assess whether an informed teacher ("Teacher") can act according to his learner's false belief and possibly even eliminate it. This further evaluation of a supervised training model functions to test the hypothesis that the a Teacher and Learner possess similar neural responses at the time of prediction (of the other's choice) and at the time of choice. If the Teacher is engaging in proper social reasoning processes, it is expected that they will engage in cognitively effortful processes in an attempt to "think like the learner".

Through the following analysis, we aim to continue the non-spectatorial work on valuation systems involved in successful social interaction, specifically cooperation. Naturalistic[2] dual/dyadic EEG recording allows for confidence in our temporal accuracy, so we will use it to establish a predictive measure to determine whether a participant is currently in a prediction or choice state during a FBT. The scarcity of research on valuational decision making within the context of social interactions has led to our interest on mental and physical coordination, beyond the joint action approach to cooperative success.

Interactive False Belief Teacher-Learner Task

Payoffs: 2 participants, A and B, form a dyad and each alternates being a Teacher and Learner. Each participant chooses between a Left and Right button. Each button is probabilistically associated with either a high (10¢) or a low (5¢) outcome. If A and B choose the same button, they receiv a 10-fold bonus in the payoffs. Results analyze data from one dyad.

Task: On each trial, each participant indicates their prediction about the other participant's choice, and then indicates their own choice. Every 3-4 trials, a reversal signal is shown to the Learner, who now becomes the Teacher. The Teacher switches her choice to the now low (5¢) outcome (Fig. 1, B, while maintaining the same prediction for the Learner (Fig. 1, A) because the Learner does not yet know that the reward probabilities have swapped. Over the next 2-3 trials after a reversal, the Learner gradually switches her choice to his now optimal outcome (Fig 1, B. The Teacher adjusts her prediction of the Learner's choice in approximately the same degree (Fig 1, top left). Similarly, upon receiving the reversal signal Teacher's reaction time for her prediction of the Learner dramatically increases (Fig 1, C) – despite the fact that she maintains her prediction from previous trials (Fig 1, A). This suggests that the Teacher engages in cognitively effortful ToM processes to ensure that they earn 50 points in each trial as the Teacher waits for the Learner to abide by the new reward probabilities.

Example: Suppose that in trial X, Left is associated with a high outcome and Right is associated with a low outcome. Teacher$_X$ and Learner$_X$ have both learned the reward probability and thus both pick Left (10¢) to maximize their payoff. Then, Learner$_X$ receives the reverse signal and becomes Teacher$_{X+1}$ and Teacher$_X$ becomes Learner$_{X+1}$. Even though Teacher$_{X+1}$ knows that Right has the high outcome, Teacher$_{X+1}$ (1) predicts that Learner$_{x+1}$ will choose Left and (2) chooses Left (L does not know the reward probabilities have reversed) so they can collectively gain 50 points.

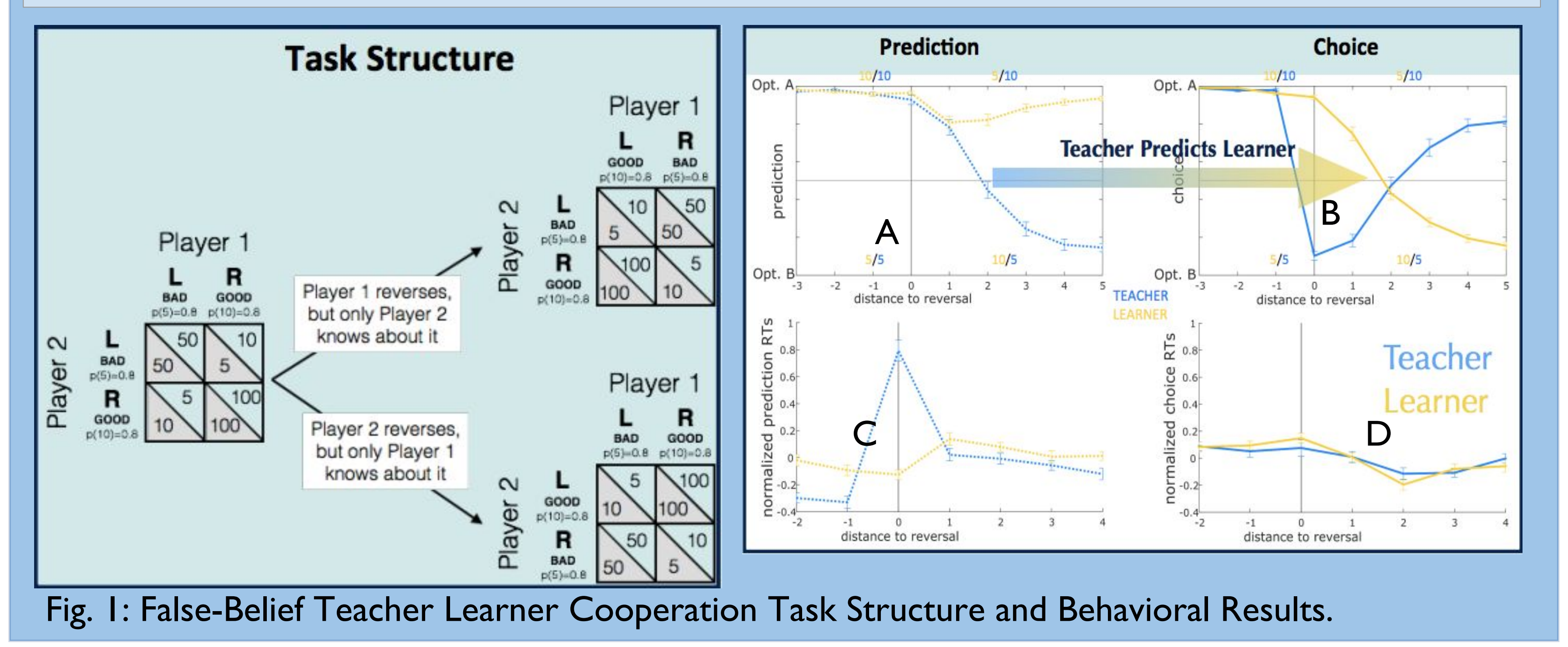

Fig. 1: False-Belief Teacher Learner Cooperation Task Structure and Behavioral Results.

Approach

EEG Data Collection

Whole-brain imaging data were collected using a dyadic 2x128 electrode BioSemi ActiveTwo channel system. This allowed for temporally-linked events to be time locked to one computer and synchronized online.

Preprocessing, as well as ERSP and ITC identification, were done using EEGLAB in Matlab. First, data were sampled at 1024 Hz, followed by sufficient "cleaning" for analysis, including rejection of bridged electrodes, filtering of high and low data, removal of gross and fine artifact, and removal of line noise. Second, interpolation of channels, re-referencing to the average electrode, adaptive mixture independent component analysis (Delorme et al., 2011), and dipole fitting were done to eliminate a wide variety of artifacts and localization of independent components using a BEM head model (DIPFIT; Delorme et al., 2012). Two datasets were created and independently epoched for either the ninety-three Prediction or Choice events. From this, the ERSPs and ITCs were visually inspected for frequencies and times of interest to be used in the classification process to determine whether a participant was currently predicting their partner's choice or choosing for themselves.

By careful feature selection, the original variables of the data were reduced in dimensionality to the point where further analysis and exploration could take place. These principal variables, while only a subset of the original data, preserved the structure of the data of interest. From the EEG ICA activiations, there were three dimensions: Signal x Time x Event. Production of these maximally temporally independent source signals available in the channel data allowed for activation localization and selection for those which account for significant variance. Entry into the time-frequency domain then produced a four-dimensional matrix: Frequency x Time x Event x Signal, which form the basis of our two-dimensional feature matrix (Event x Feature (Frequency) x Time Bin).

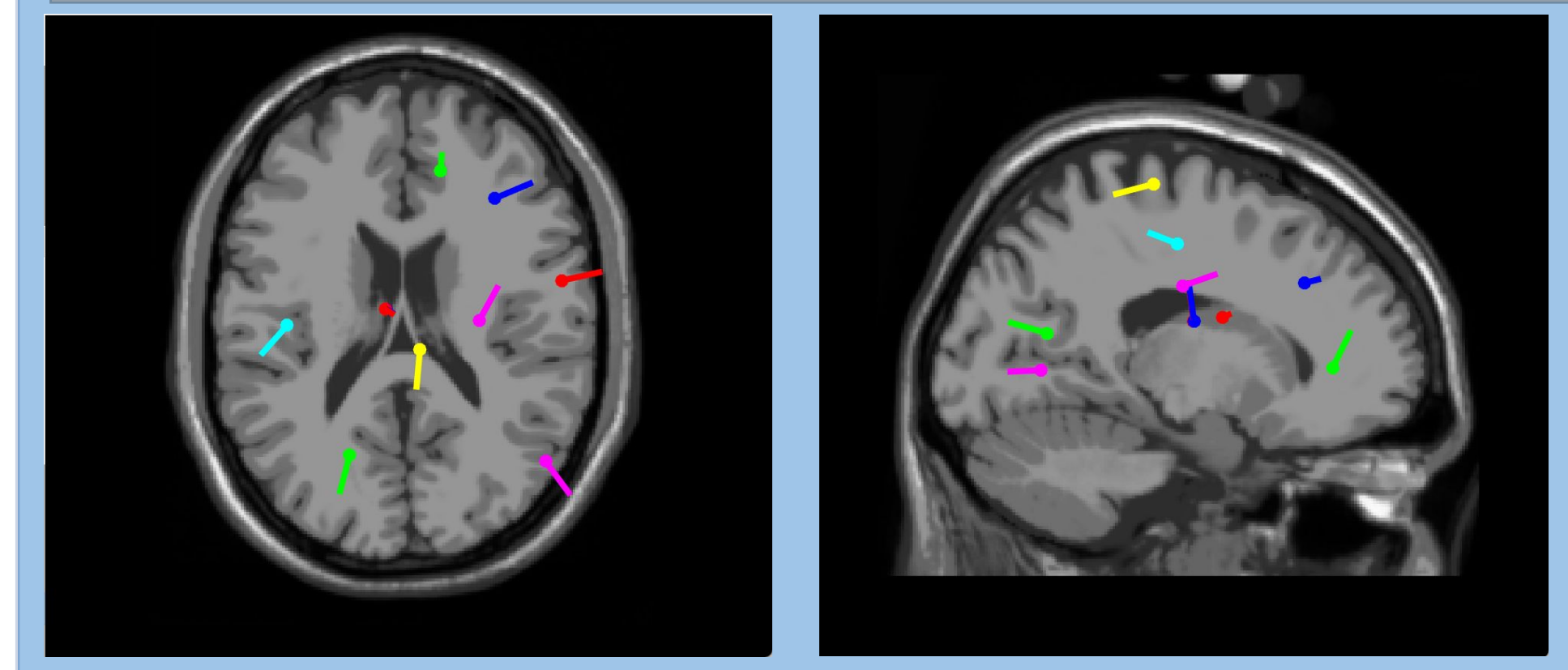

Fig. 2: Component Dipoles for the 10 ICAs of interest.

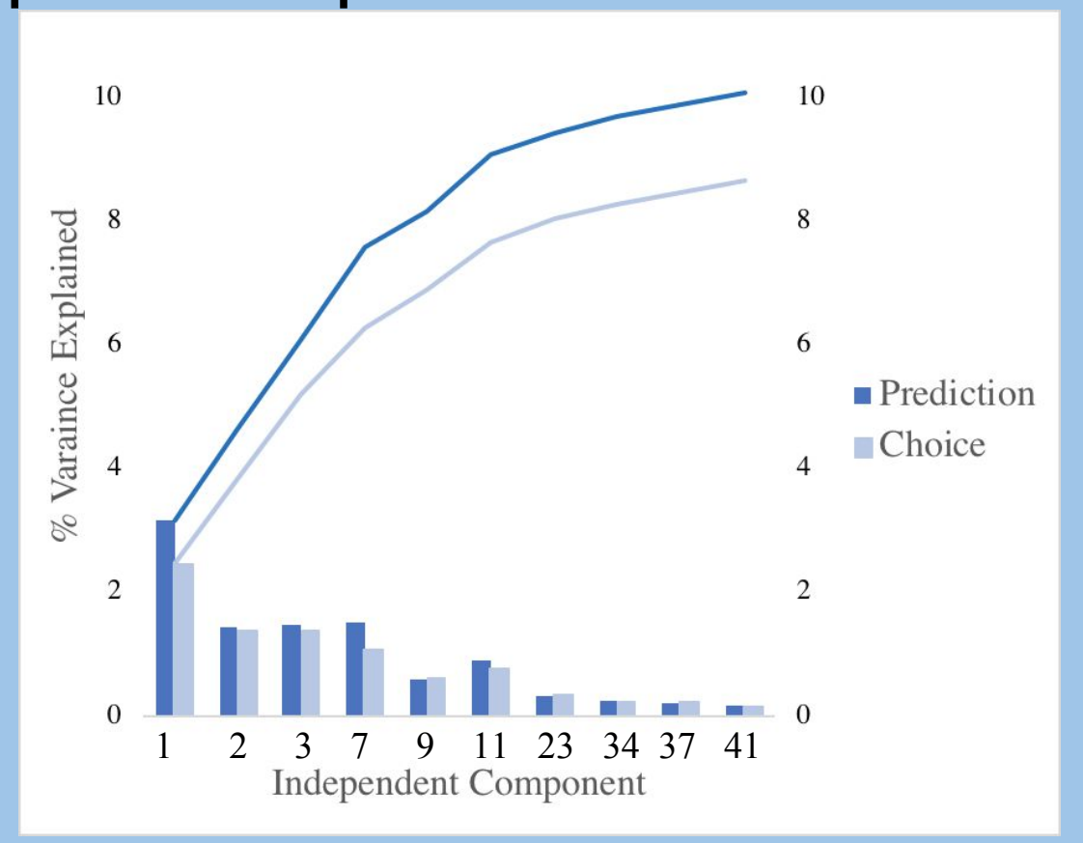

Fig. 3: Comparison of Variance Attributed to 10 Most Significant Independent Components.

Analysis via Machine Learning

Our feature matrix was analyzed in Python using scikit-learn. Three machine learning techniques were trained and evaluated: a Logistic Regression (LR), a linear Support Vector Machine (SVM), and a nonlinear SVM using a radial basis function kernel. Each model was tested both with and without using principal components analysis for dimensionality reduction, during which we used each component preserving > 1% of the data variance (Fig. 3). For LR, we tested various proportions of a test/train split, the best performing of which are reported in Table 1. For the SVM analyses, we used a Leave One Group Out approach, holding back ~10% of the data on each of 10 iterations for testing. Additionally, we investigated which range of frequency bands carried the most useful information for classification by running these analyses on both high and low bands, the results of which are also reported alongside the all-band analysis in Table 1.

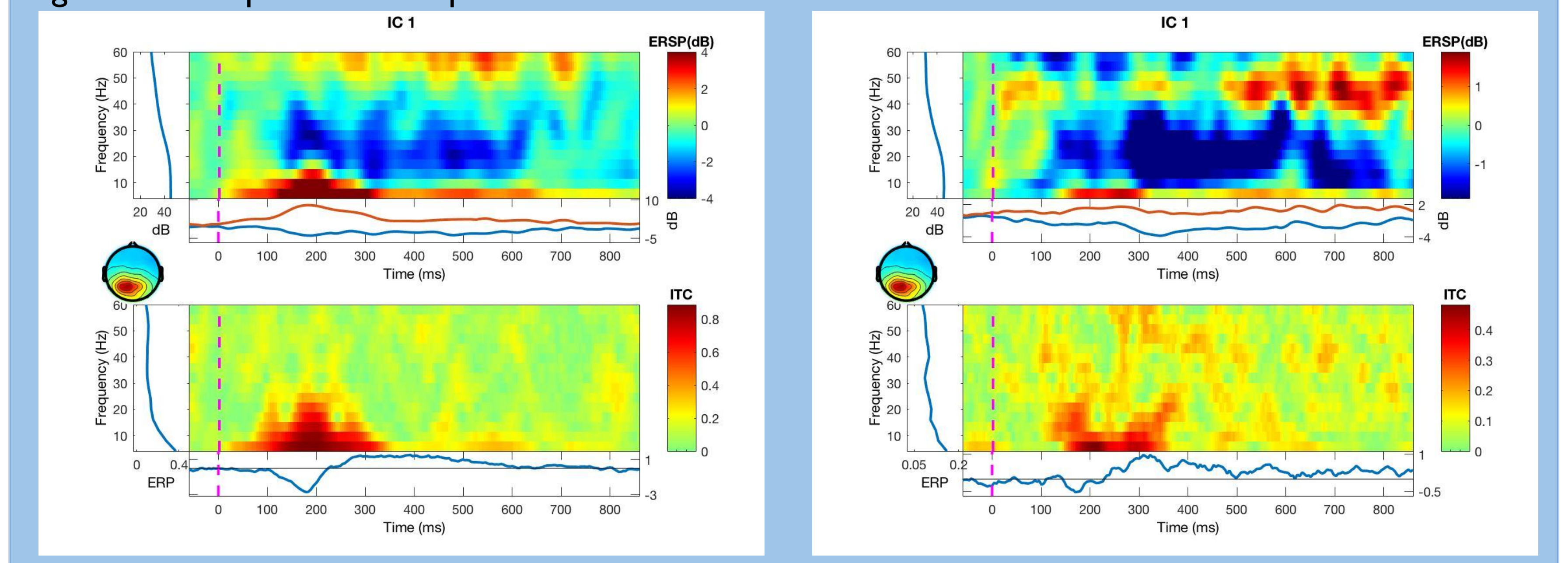

Fig. 4: Prediction and Choice event-related spectral perturbation (ERSP) and inter-trial coherence (ITC) in component 1.

Evaluating Classifier Performance

Frequency Bands	Logistic Regression (PCA)	Logistic Regression (no PCA)	Linear SVM (PCA)	Linear SVM (no PCA)	Nonlinear SVM (PCA)	Nonlinear SVM (no PCA)
All	*0.789473684*	0.8	0.76166667	0.809444444	0.645	0.787777778
Low (5-30 Hz)	0.746666667	*0.789473684*	0.73055556	0.793333333	0.698888889	0.793888889
High (30-60 Hz)	0.733333333	*0.736842105*	0.69166667	0.717222222	0.637777778	0.678333333

Table 1: Mean classification accuracies for the models used, for both the centered and scaled (no PCA) and dimensionality reduced (PCA) data. The low and high frequency band splits resulted in equal numbers of data points. For the Logistic Regression results, those in italics reflect an optimal test/train split of .2/.8 and the others a .4/.6 split.

Discussion

Model Comparison: Each of our three classification models reached peak performance around 80% successful classification (i.e. prediction or choice state) on the test data. Although there were small differences in performance, they are not large enough to declare one model clearly superior in parsing this data.

Principal Component Analysis (PCA) Performance: One surprising result is the relatively worse performance each model, especially the SVMs, exhibited when using the PCA transformed data compared to the non-dimensionality reduced input data from our putative feature matrix. The reduction in this case was from 320 features per observation to between 35 and 40 principal component features. It could be that this relatively modest change, or perhaps the lack of correlations between our selected signals, accounts for PCA not being a fruitful approach here.

Frequency Band Comparison: For each of our models, restricting the input data to only the low frequency bands of interest (approximately theta, alpha, and beta) either essentially preserved model performance, or improved it in the case of nonlinear SVM. On the other hand, only analyzing the higher frequencies in our data set (approximately gamma waves) consistently reduced model performance, by an average of ~6.6%. This is strong evidence that the most relevant neural code necessary for distinguishing between prediction and choice is in lower frequencies.

Future Directions: In this exploratory project we looked at one member of a participant dyad to see whether the differences in neural responses differed consistently enough to classify between a choice and prediction event. Examining the data further for false positives and negatives would allow for better understanding of the real accuracy of the classifier. Using this approach, we could expand this project to the full range of dyads to train and classify between choice and prediction across a sample group. Perhaps by increasing the specificity of both time and frequency bins further classification could distinguish between choice and prediction in respective teacher and learner roles and identify the choices made by participants based on their EEG data. Further, since the data is synchronized online, it would be possible to design interactive brain computer interface tasks, which would not only reduce motion noise but also provide a closer reading of the partners perspective. Interactive BCI task would then be able to rapidly transform these neural signal readings and portray it as movement on the screen, enabling new avenues for communication between participants.

Acknowledgments

We gratefully acknowledge Michael Spezio, Yuqing Lei, Vanessa Hayes, Tessa Rusch, and Jan Glaescher for their prior work in the data collection and analysis with focus on uncovering patterns of theory of mind and event related potentials.

References

1. Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic Theory of Mind processing in adults. *Cognition*, *162*, 27–31. https://doi.org/10.1016/j.cognition.2017.01.018
2. Turner, R., & Felisberti, F. M. (2017). Measuring Mindreading: A Review of Behavioral Approaches to Testing Cognitive and Affective Mental State Attribution in Neurologically Typical Adults. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00047
3. Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A Second Look at Automatic Theory of Mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, *26*(9), 1353–1367. https://doi.org/10.1177/0956797614558717