Comparison of Machine Learning and Human Decision Making in Noisy Datasets

S. NGUYEN, R. SRINIVASAN;
Dept. of Cognitive Sci., Univ. of California Irvine, Irvine, CA

Abstract

The use of Artificial Intelligence (AI) across a variety of fields has rapidly become commonplace. Often AI is tasked with handling tasks or information that may be vast and inefficient for a human to handle. In turn humans, when tasked with working with these systems, often default to the decisions of the AI. However, as the application of AI grows beyond this and AI is integrated into environments or frameworks that are not optimal for it, a new problem may begin to arise. When the AI and a human are inaccurate in a task or situation, how can they work together to improve? The research conducted in this paper seeks to tackle this growing possibility. Through the observation of subject's and machine learning in a problem where both perform with poor accuracy, the possible differences between humans and machine learning in certain tasks can be observed. Through a combination of subject explicit information weighing, machine learning prediction of self and humans information weighing, and neural activity recorded via EEG a more complete picture of human/artificial intelligence decision making can be resolved. In the scope of this study, it was recognized that in subject and machine learning pairings there are differences in how information is weighed, which translates to different sets of correct responses. However, the expertise of the subjects varied. This was informed through the explicit information selection and the machine learning replication of the subject's behavior.

Introduction

In the modern world, technology has increased its integration into everyday life, it is common to see artificial intelligence (AI) take on tasks of varying complexities and objectives. In context, AI is the replication of human intelligence. An AI system would be thought of as a system which reacts to stimulation or stimuli in a way which would be considered similar to humans (West, 2022). This goal of artificial intelligence is based on the idea that human action and reaction is a system of models which interpret and learn from experiences to inform future decisions (Ma, 2017). Machine learning, as a subset of AI, is considered a system which can similarly learn from experiences or data and produce a response which is congruent with what is interpreted from the data (Brown, 2021). It is becoming common practice to integrate various AI and machine learning to everyday tasks for the purpose of either assisting humans or replacing the need for humans altogether (West, 2022). However, the specialization of machine learning can also pose a problem. Overspecialization moves away from generalizability; producing an AI which can only perform decently across various different situations, generalization in this usage

refers to the application of AI systems to a variety of complex situations and tasks. For the time being, AI system development is led by machine learning, which inherently weakens the generalizability of AI as machine learning continues to overspecialize (Brown, 2021). Overspecialization in this context refers to the specification of a system to a particular task to the point where it hinders or completely removes the ability for that system to be applied in more general or uncertain situations (Badman, 2020). In recent years, ChatGPT, an artificial intelligence chatbot which is trained to respond and converse in a human-esque way, has been featured in news for its ability to trivialize human tasks ranging from daunting to insignificant (OpenAI, 2022; Mollick, 2022). However, despite this innovation it can also produce meaningless or nonsensical text and is prone to the inability of making sense of simple misinformation (OpenAI, 2022). This artificial intelligence highlights many of the concerns of the need to generalize AI systems as well as integrate these systems so that humans and AI can assist each other.

Artificial intelligence is not at the point of innovation where it can be integrated into a complex system without the oversight of a human. In fact, current trends in artificial intelligence (AI) decision making research look at the augmentation of human decision making when assisted with AI (Steyvers, 2022; Zhang, 2020). This work deals mainly with the question of trust from humans towards the AI. It is an expectation that the AI will outperform the human in decision making and therefore, it is up to the human to develop some trust in the AI's decision making and place a lot of weight on its decisions, while still verifying the validity of the AI's decision. Therefore, it can be implied that the study of decision making further delves into the development of human mental models which take into account the AI agent.

However, there are limitations to this approach in the study of decision making. First, it should be considered that there are, and can be, tasks where the human and the AI perform similarly on average, but vary on the subset of instances they get right, such as in the case of certain visual noise tasks, where humans and AI are not aligned in their error boundaries (humans are correct when AIs are not and AIs are correct when humans are not) (Steyvers, 2022). Second, it is a natural step in the field to generalize the task to a more real world task, where humans have lifetime experiences that provide contexts to decision making. These lifetime experiences allow humans to infer information, either explicitly or implicitly, not directly provided in the inputs. Such human "intuition" can potentially improve performance in tasks but can also inhibit performance by introducing biases into decision making. Often a study will look at a task where the decisions made are not generalizable, which has its own merit in the field of study. However, tasks which further mimic real world tasks allow for more meaningful analysis.

Previous Study

Zhang et al. 2020, observed human decision making in conjunction with machine learning assistance in an income prediction task. The experiment tasked individual's with predicting income above or below a threshold using census information provided via the 1994

Census Data published as the Adult Data Set in University of California Irvine (UCI) Machine Learning Repository. The census dataset used consisted of 48,842 instances of surveyed persons, each described by 14 attributes. Zhang et al. reduced the amount of provided information to 9 attributes: age, sex, race, marital status, years of education, workclass, occupation, and hours per week. Each attribute and value was followed by a chance value (indicating the likelihood that an individual with the corresponding value would make more than $50,000 (Figure 1).

Attributes	Value	Chance
Age	50	4
Sex	Male	3
Race	White	3
Marital Status	Married, spouse civilian	4
Years of Education	10	2
Workclass	Private	2
Occupation	Executive & managerial	5
Hours per week	40	2

Figure 1: Sample screenshot from Zhang et al. income prediction task. The chance value was used in the original experiment to assist in guiding participants. In the experiment conducted in this paper, the chance value was omitted to allow for more flexibility in participants forming individual approaches to the experiment.

The design of the experiment consisted of two groups. Both groups were provided with feedback informing them of their answer and the AI's answer in comparison to the correct answer; and one group also received the confidence level of the AI in its answer. When placed in test trials, participants were allowed to switch answers after seeing the AI's answer. The design of this research was intended to study the calibration of trust in AI as a function of confidence. The design of this experiment was adequate in allowing researchers to study the calibration of trust with the AI and the effect of AI confidence on the calibration. However, the design was limited by two factors: 1) the humans were not able to perform better than the AI in the dataset and 2) the human and AI were largely aligned in their error boundaries (i.e. when the AI was wrong the human would likely be wrong as well) (Zhang, 2020). Steyvers' 2022 study examined the interactions between human and AI agents using various machine learning classifiers as well as several combinations of human(s) and said machine learning classifiers. Using visual stimuli with noise, a task was devised where humans and AI performed well, but on different noisy

versions of the task. This concept of similar levels of performance, but in different instances highlights a difference in the way information is taken in and weighed. In a multi-feature decision making task, such as in Zhang's paper, the features which carry more weight for humans and for AI is not as explicit, but it is possible that there is still a weighing of information that humans and AI differ on, and therefore, certain subsets of the task where they perform differently.

From the interpretation of these studies several questions come to mind. First, does the mental model that the human uses weight information in the same manner as the AI? If there are differences in the models, do they translate to differences in which instances humans and AI's succeed? Would these differences be evident in electroencephalograph (EEG) signals as information is presented? A thorough understanding of these differences would reveal more information about the faults in either human or artificial intelligence and could be applied toward the integration of more efficient human and AI systems.

Application of EEG

Previous research using EEG devices has looked at the link between decision making and event related potentials (ERP) in the brain (Si, 2020; Wang, 2015). More specifically, a spike in the amplitude at around 300ms (p300) following the presentation of highly salient information (a cue relevant in making a decision), is regarded as an important marker in the study of human decision making. The amplitude or magnitude of the waveform is positively correlated with the information's overall importance toward making a decision, a greater magnitude indicates a greater shift towards making a decision (Mansor, 2021). Furthermore, the discrimination of salient information and the associated activity in the brain, specifically p300 ERPs has already been previously applied such as p300 Speller Brain-Computer Interfaces (BCIs) (Kirasirova, 2020; Won, 2022). Subjects are shown an array of letters and when the letter they are in search of (such as to spell a word) begins to flash the subject will have a greater cortical response to the letter than to the flashing of an unimportant/unused letter. Conceptually, this independent metric of information salience can naturally be applied to multi-feature decision making. In many situations, the importance of one piece of information will be more valuable and therefore, an individual may be more attentive to said information which should be reflected in their cortical response.

Proposed Design

The proposed experiment is a modified version of Zhang's binary income prediction task that seeks to address the concerns and ideas brought forth by the aforementioned research. Rather than seeking to measure trust however, this task was performed to study the weighing of information in human's vs. machine learning. The experiment used a modified version of the same income prediction task, however, the focus was shifted away from human calibration of trust in AI. Nearly all the same features were used, however, rather than years of education, the highest achieved degree was used. It was highly likely from the dataset that these two values

held similar meaning and for a human the highest level of education would be more rapidly processed than the numerical value of years of education. Unlike the aforementioned prediction task, individual's were not shown any information regarding the machine learning's decisions and were not shown "chance" values. The omission of the chance value was to make the information presented to the individual and to the machine learning congruent; the removal of the chance value also allowed more flexibility for the subject to navigate the information in a natural way. Due to the aim of this experiment, it was imperative that the conditions of the subject and the machine learning were kept relatively congruent. To further deviate, rather than presenting all information to the individual at once, each feature was presented for a short period one after another (i.e. a steady stream of information). By isolating the presentation of each feature and observing the neural reaction to said feature, we aimed to obtain an independent measure of the importance of the feature. This importance metric is based on the p300 marker of stimulus salience (Mansor, 2021; Si, 2020). Furthermore, by applying a predictive Machine Learning model, the model was expected to be able to mimic the human's performance and make an estimate of the feature weight in the human decision making process. Finally, the feature importance weights as determined by eeg recordings and machine learning classification were compared to the explicit features an individual chooses to be given when presented with an incomplete selection of information.

Methods

Participants

IRB approval was obtained prior to any data collection. This study looked at 5 participants. No information that may be personally identifying was collected over the course of the study. Subjects were informed of the usage of an electroencephalogram (EEG) device for 30-40 minutes of the session. All participants were informed that they would be compensated $30 per session across 2 sessions for a total of $60. Furthermore, subjects were informed of the option to end the experiment early with a payment appropriate to the amount of time spent in the experiment prior to early release. All subjects consented to participate in the study prior to starting the first session and after being briefed of the requirements, structure, and parameters of the experiment.

Overview

Participants were tasked with predicting whether an individual who is described from a set of demographic descriptors would make above or below the U.S. median income (~ $50,000). The demographic information provided for the tasks was (n = 9):

- Age
- Occupation
- Hours per week
- Workclass
- Marital Status
- Race
- Highest Degree
- Gender
- Native Country

The method of providing information was divided into two distinct tasks, a “full feature” and a “feature selection” task. In the full feature, participants were presented a randomized sequence of all nine of the demographic descriptors before making a prediction. In the feature selection, participants were provided with a combination of three features and asked to select one more feature to be presented before making a prediction.

The experiment was conducted across two sessions for each subject. The first session consisted of three blocks: a training block (a block congruent with the full feature task) of 100 trials, a full feature block of 195 trials divided between 3 sub-blocks, and a feature selection block of 84 trials. The second session consisted of two blocks: a full feature block of 240 trials divided between 3 sub-blocks, and a feature selection block of 84 trials. All subjects were given feature sets pulled from the same modified dataset. The randomization was manipulated so that all participants would all see the same feature sets as each other in the training block, full feature block, and the feature select block. Participants were attached to a 128 channel EEG during both full feature set blocks. EEG samples were recorded at 1000 Hz.

Stimulus Dataset

The dataset used was a modified version of the Adult Learning Set available on the UCI machine learning repository. The original dataset presented 49,531 entries of demographic combinations with the associated income value. Income value was adjusted to reflect medium incomes in the current year (2022). The initial dataset was tested and run by a histogram-based gradient boosting classification tree based machine learning model on a 70/30 split (70% to train and 30% to test) which was trained and tested on the initial dataset and cross validated to verify an accuracy of 83% (Pedregosa, 2011). Prior to producing the experimental set, a model was trained and tested on an intermediate set which removed all demographic information except for the nine expected to be used in the experimental dataset. Features were removed from the final dataset to remove redundancy (such as years of education vs. highest degree) or values which are less commonly understood (such as capital gain, capital loss, and relationship). The initial dataset was also cleaned to remove any entries where any attribute value was absent. The resulting intermediate dataset of 45,849 entries when trained and tested by a machine learning model under the same parameters was verified to have an accuracy of 80%. Following the initial and intermediate verification, feature sets were selected from the intermediate dataset to produce a testing set (435 feature sets) and training set (100 feature sets) which both represent the original dataset value representation (through comparing percentage representation for each value) and provide a dataset which the Machine Learning Model operates at a 60% accuracy. The accuracy was confirmed by running and cross validating a trained model on the experimental dataset prepared for the experiment and the machine learning was tuned to maximize operation on said dataset, the same tuning parameters were used on subject performance analysis. Furthermore, the experiment dataset was selected so that 50% of the cases were individuals who make less than $50,000 and 50% of the cases were individuals who make more than $50,000. The same

experiment dataset was used for all subjects and distributed across tasks so that all full feature task sets would be the same for all subjects (randomized order of sets and feature presentation) and all feature select sets would be the same for all subjects.

Training and Full Feature Decision

The training block was the same as the full feature block, with the exception of no EEG attached during the training period. The training block acted as an aid for participants to get a better understanding of the task and feature values before starting the experiment with the EEG attached. In the full feature block, participants were presented with 9 features: Workclass, Highest Degree, Marital Status, Race, Gender, Native Country, Occupation, Hours per Week, and Age. Features were presented in a randomized sequence at the center of the screen for 500ms followed by 300ms of a blank screen. Following the presentation of all information, the participant was given a 1000ms response window and was asked to make a prediction of whether an individual with all the presented features would make greater than $50,000 or less than $50,000 (Figure 3). Following each prediction, participants were provided with feedback informing them whether their answer was correct or incorrect. The feedback was presented on the screen for 500ms. After the feedback presentation, a blank screen was presented between trials for 500ms to 1000ms. A fixation was presented in the center of the screen where the features would be presented for 500ms to 1000ms before each trial. Participants were instructed to consider their answer as more information was revealed with the understanding that each trial has a limited response window following the presentation of all information.

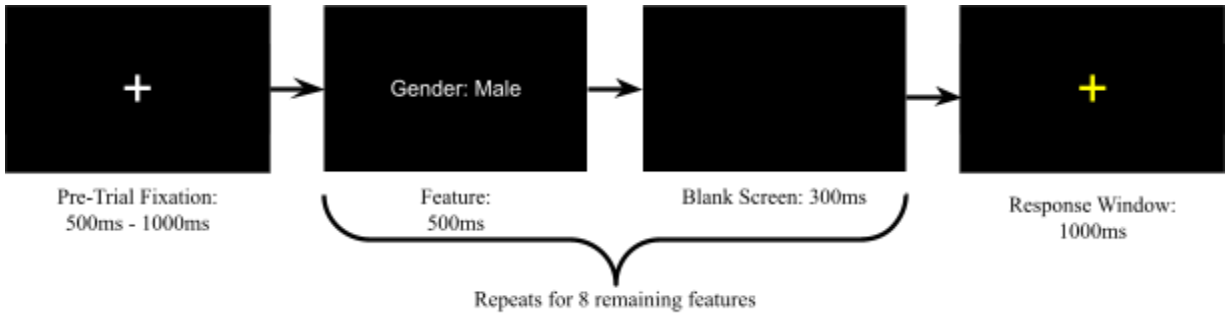

Figure 3: Example of the Full Feature Decision trial. All nine features are presented separately and the onset is separated by a blank screen. All trials start with a white fixation and the response window is indicated by a yellow fixation.

Feature Selection

The feature selection block consisted of the participant being provided an incomplete set of information. Participants were provided with 3 of the 9 features and the associated value. Before making a final prediction, participants were asked to choose one feature of the remaining 6 features to add to the known set of features (Figure 4). The combination of features were determined so that in one session a participant would be exposed to all possible combinations of 3 features once.

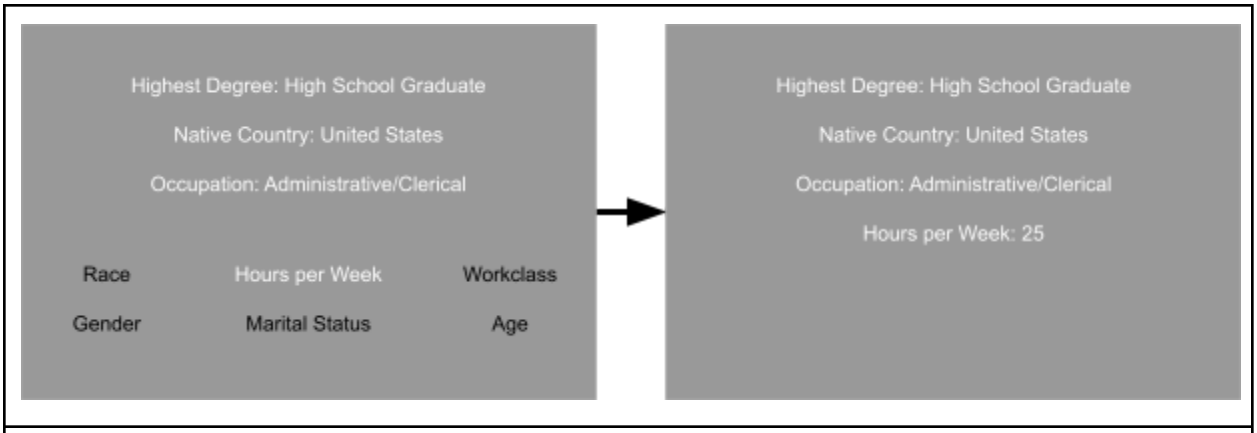

Figure 4: Sample frames of Feature Select screen. Participants are shown three features and select a fourth feature from the bottom six (current selection is highlighted white) (left frame). Participants are shown their selected features and are required to make a prediction (right frame).

Data Processing

The accuracy of the subject was recorded and compared to the machine learning model. Specifically, the similarities between the subject and machine learning were captured as accuracies when aggregating the correct decisions between the subject and the model (the maximum possible correct responses if either agent answered correctly), and differences were observed as the percentage of the correct responses where only the subject was right or only the machine learning was right.

For the duration of the Full Feature Decision Task, all participants wore a 128 channel EEG cap and were recorded in sub-block separated recordings (one recording per sub-block). Following data collection, all EEG data from the participants was processed. Data was run through a 1Hz to 50Hz bandpass filter and an independent component analysis (ICA) to remove eye blinks and other body movements which may interfere with the EEG readings. The ICA used the three frontal channels (Fp1, Fp2, and Fpz) situated above the eyes to measure EEG signal eye movement artifact correlations. EEG recordings were segmented by feature/item onset (9 feature segments and the response cue for a total of 10 segments) where each segment consisted of 200ms prior to stimulus onset and 800ms after the initial onset. In terms of the feature presentation, this window following onset corresponded to the 500ms of the feature presentation and the 300ms blank screen that followed.

After initial EEG processing, individual trials were omitted from EEG analysis if the participant failed to respond in the response window. In the cases that the subject responded, trial segments were excluded if the EEG recording was greater than 2.5 standard deviations from the median of their trial segments. EEG channels were omitted from subject EEG analysis if the channel was unresponsive during the experiment. Channels that were responsive, but recorded a mean activity greater than 20 standard deviations above the median, were incorporated in ICA analysis to correlate with properly functioning channels. The trials which were removed for EEG analysis were not removed from the subject's behavioral analysis.

Each trial was segmented so that all feature onsets could be extracted separately. The resulting segments were categorized by the appropriate feature that was presented and averaged to produce an average EEG segment for each feature. EEG recordings for each segment of a trial were categorized and averaged by each respective feature. Initial observation of EEG activity between onset and the next feature or trial end was collected for observation. After filtering and eyeblink removal of the EEG, the EEG segments were baseline corrected and average referenced in order to further remove noise and clean the EEG signal. Following this, a singular value decomposition (SVD) analysis was performed in a window of 250ms - 400ms after item onset to produce single EEG waveforms for each feature per subject. This was performed in order to further extract the p300 signal. Through this, a direct comparison of p300 amplitudes within the 250ms - 400ms window could be performed within and between subjects.

Analysis via Machine Learning

The order of feature presentation as well as response for the trial was recorded and cleaned to remove trials without responses. A histogram-based gradient boosting classification tree based machine learning model used a 75/25 split to train (75% of valid trials) and test (25% of valid responses) on the subject's dataset using the subject's response as the value to predict (Pedregosa, 2011). Following the running of the model, the model was run through a partial dependence calculation in order to calculate the effect specific values within a feature category have on the ML prediction. When calculated, the partial dependence outputs a value between 0 and 1 for each value within a feature category. The importance of these values in the model's decision is relative to a "null value" or value which indicates no influence on decision. The value itself adjusts according to the split between 0 and 1 responses. For example, if a subject responds with 1 60% of the time, the null value moves to 0.6 and feature values with a partial dependence of 0.6 would not be useful in the ML's decision making tree. A value greater than the null value indicates a move of the ML toward responding with 1, while a value lower than the null value indicates a move of the ML responding with 0. For each subject, the response prediction partial dependencies were compared within features. The maximum and minimum dependencies were taken as a range (the value which moves the ML most toward 1 and toward 0 respectively). These values were collected and interpreted as feature importances by the model; the more variation within a feature, the more useful it would hypothetically be for the model. Similar partial dependencies were calculated with the model predicting the correct answer for each subjects' valid trials and averaged after verifying consistent model behavior across subject datasets. The resulting dependence values were processed similarly to the values produced by the model predicting the subjects responses. The resulting correct answer prediction partial dependence plot was compared to the subject's response prediction plot to observe differences in feature weights. To bring together these separate measures of the subject's feature weighting, the feature selection proportion, the weighted importance from the machine learning, and the magnitude of p300 responses were compared.

Results

Behaviorally, subjects had a mean accuracy of 0.56 (SD = 0.02) and median of 0.56; and the machine learning model had a mean accuracy of 0.63 (SD = 0.02) and median of 0.64, both were adjusted to remove trials where no response was recorded from the subject. When the performance of the subject and machine learning model was combined the average accuracy of one or both agents were correct was 0.81 (SD = 0.02) and median of 0.81, where 0.23 (SD = 0.02) of the correct responses were unique to human subjects and 0.30 (SD = 0.03) of the correct responses were unique to the machine learning model (Figure 5). To examine the possibility of a subset of problems which different subjects may consistently perform well on that the ML fails to answer correctly (i.e. feature sets that are uniquely interpretable by people only), the questions which subjects were correct in only were recorded and compared. The intersection of problems in which subjects were only correct was observed to be 2.5% of the average amount of subjects with only correct responses (22%).

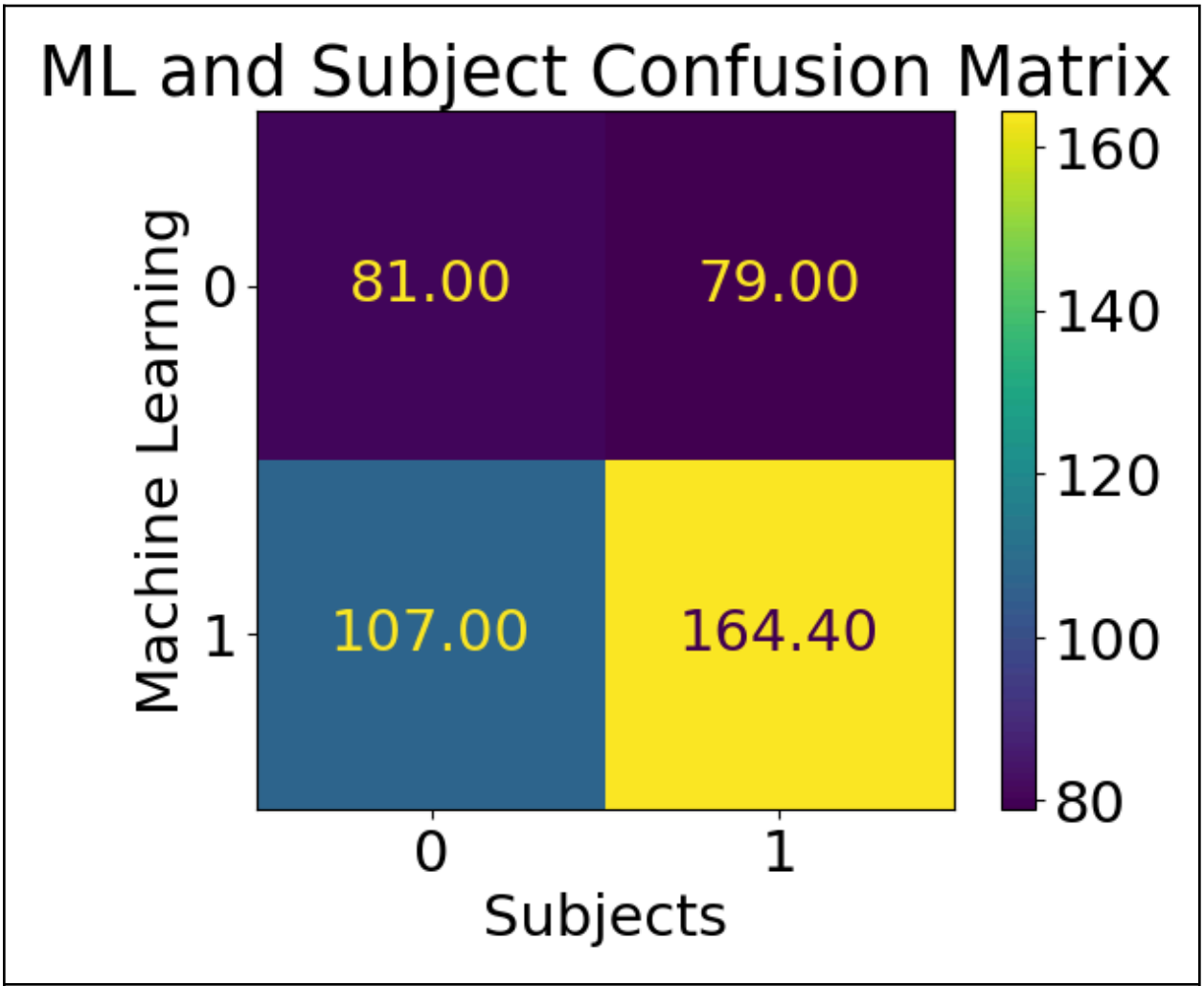

Figure 5: Matrix comparing correct and incorrect responses of ML (y-axis) and averaged subject performance (x-axis), averaging 431 trials. 1 indicates a correct response and 0 indicates an incorrect response. If both agents are in agreement and correct the x and y value component will both be 1, if the subject is correct only the x value will be 1 and the y value will be 0.

Machine Learning Prediction

The machine learning model was tasked with predicting a subject's responses using the same information that the subject was given, so that the model could report the partial dependence for each value. These values were used to determine the specific importance of each possible value in the feature category. By analyzing each feature in this way, it was possible to get a greater analysis of the importance of different values within the category type. Following this, the maximum and minimum partial dependencies within the feature were taken as a range; this range quantifies the possible "influence" a feature could have on a decision. In this sense, the greater a feature's range, the greater dependence the model has on that feature when making a decision (Figure 6).

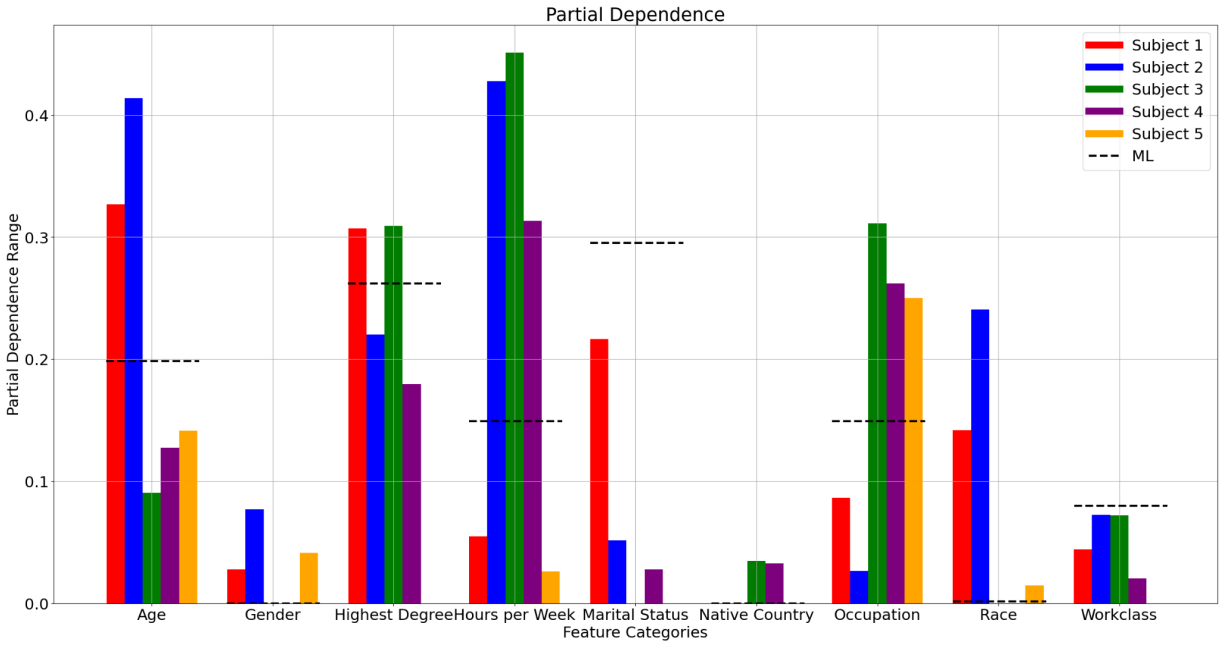

Figure 6: Partial dependencies determined by the machine learning model. The x-axis is the feature category that was used in the task. The y-axis is the range from the minimum partial dependence to the maximum partial dependence to the corresponding. The greater the range the greater possible effect of the corresponding category.

In all cases, the model was actively attempting to learn and replicate the subject's behavior. In order to verify this behavior, a receiver operating characteristic (ROC) curve was produced for each individual based on the machine learning behavior when attempting to predict the subject's responses. This was done in order to verify the machine learning model was not operating on chance, and was actively discriminating between cases where the answer was above or below $50,000 (Figure 7).

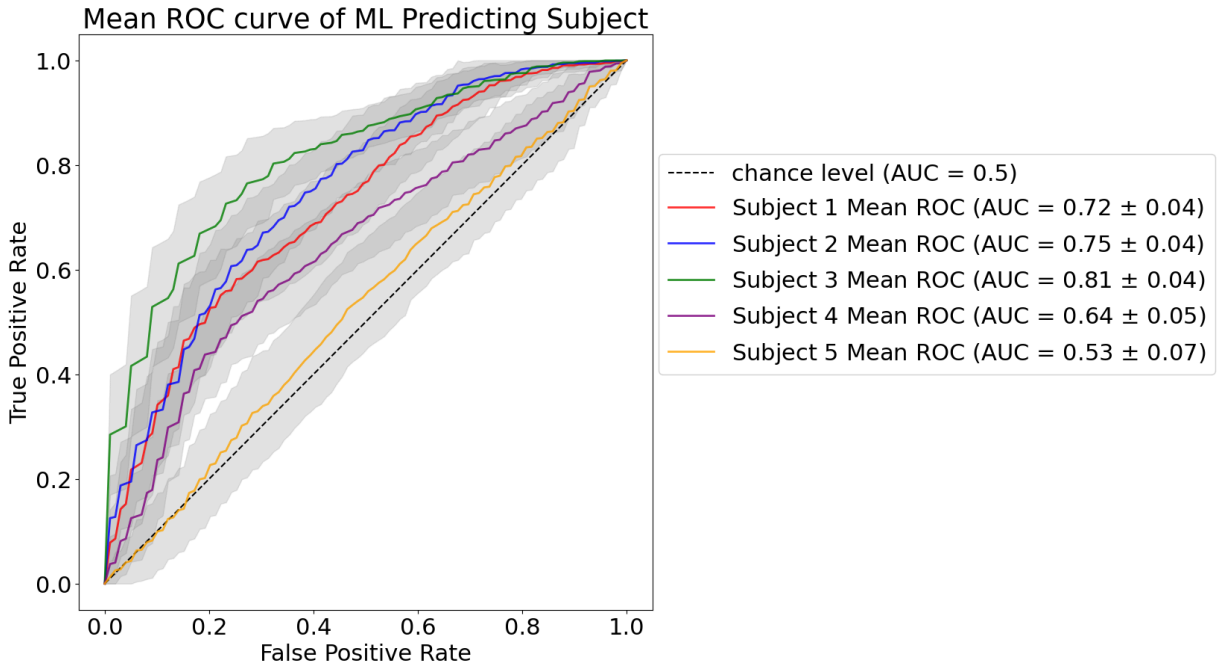

Figure 7: Receiver Operating Characteristic (ROC) curve depicting the ability of a machine learning model in replicating the behavior of a subject. The x-axis is the false positive rate and the y-axis is the true positive rate. The area under the curve corresponds to the ability of the model in distinguishing between cases where the income was greater than or less than $50,000. A greater value can be evaluated as a greater probability, with 0.50 indicating a chance probability in this dataset. All subjects had varying levels to which machine learning could replicate their responses. Machine learning was not able to produce a stable representation of Subject 5.

The performance of the ML and subjects on the same set was visualized via confusion matrices in order to verify a level of problem discrimination across subjects and ML (Figure 8). The average true negative rate (TNR) of humans was 0.56 (STD: 0.02) and of ML was 0.62 (STD: 0.02), while the true positive rate (TPR) of humans was 0.59 (STD: 0.02) and for the ML was 0.64 (STD: 0.03).

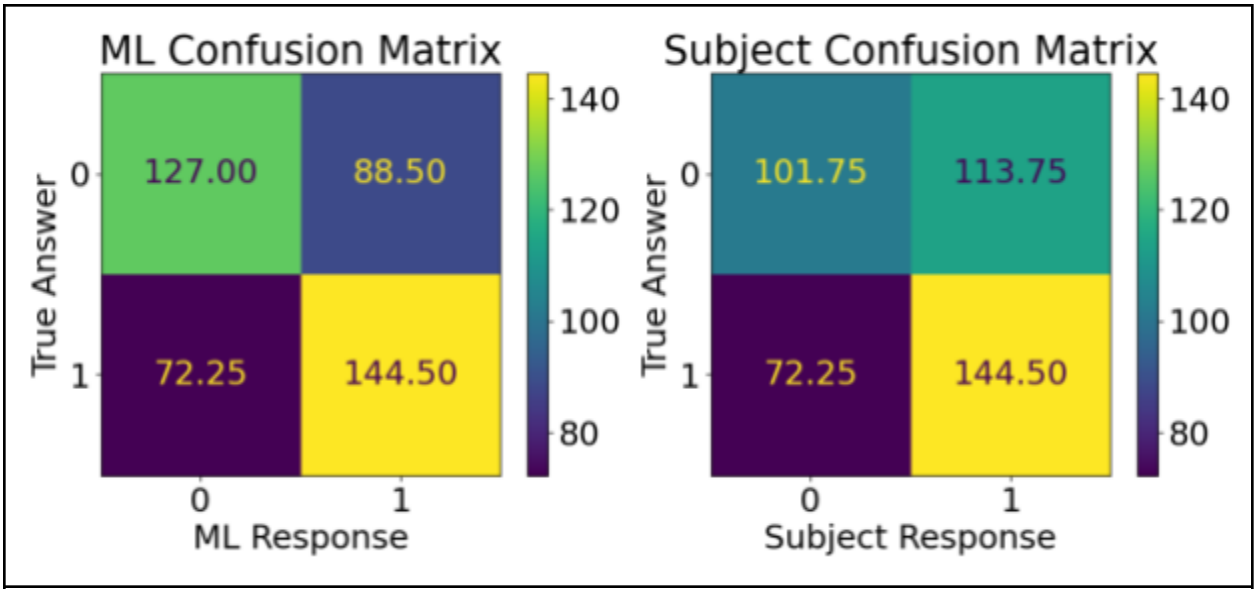

Figure 8: Confusion matrices of aggregated subjects and ML. 0 indicates a "< $50,000" response and a 1 indicates a "> $50,000" response. A true positive (TP) is indicated by a correct prediction that an individual makes "> $50,000". A true negative (TN) is indicated by a correct prediction that an individual makes "< $50,000". A false positive (FP) is indicated by an incorrect prediction that an individual makes "> $50,000". A false negative (FN) is indicated by an incorrect prediction that an individual makes "< $50,000".

Explicit Feature Selection

In addition, the features selected during the "feature select" block were recorded. Participant's chose features according to the importance that feature had in making a final decision. This process of explicitly selecting a single piece of information, reflects the individual's personal decision weights (Figure 9). The proportion of features were directly compared across subjects to observe differences and similarities in how individuals choose information, the presentation of three features also allowed for individual's to consider how information may affect the next choice, and therefore, allowing a more complex idea of how the final feature was chosen.

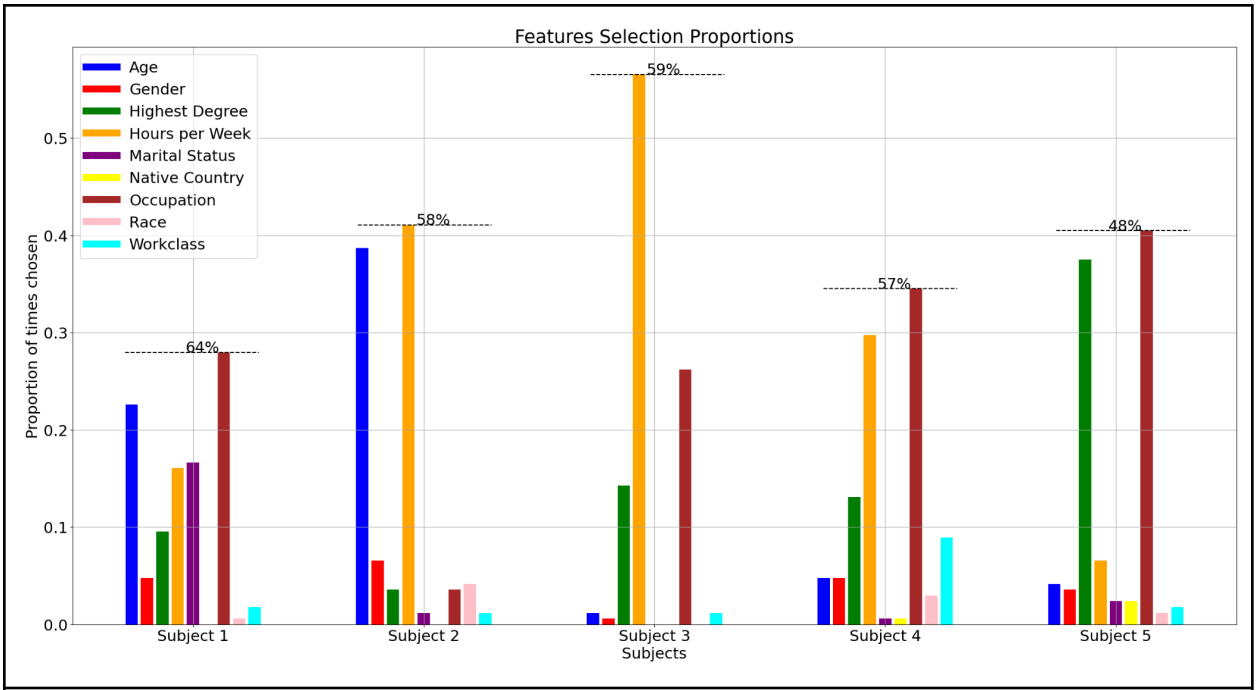

Figure 9: Bar charts of the proportion of features selected by subject out of 184 trials. The y-axis is the proportion a feature was chosen relative to the possible amount of times it could be chosen by an individual. Each bar grouping represents a subject and each bar corresponds to a feature (refer to legend for feature-color grouping). Each subject was exposed to the same set of feature combinations, but the value sets were randomized. The value above the subject's bar group indicates the subject's accuracy in the feature selection task.

Following the cleaning and segmentation of the EEG data, a singular value decomposition (SVD) was applied to the EEG data for each feature to capture the magnitudes of the p300s for each subject (Figure 10). The plotted p300s represent the average magnitude of response for each feature across both sessions, where bad trials were excluded according to the previously reported data cleaning system.

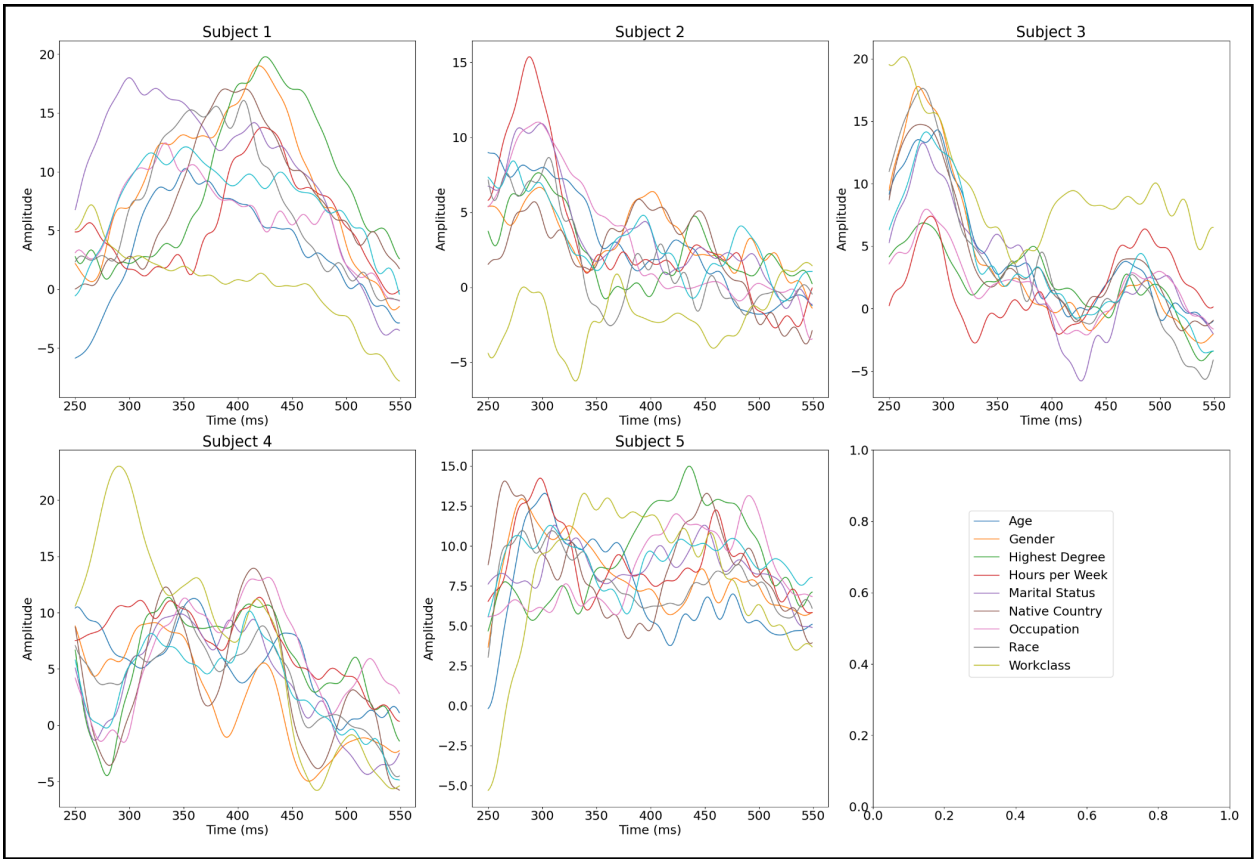

Figure 10: Plots comparing amplitude of ERPs from features for each subject across a time window where the p300 event occurs. The ERPs were plotted from 250ms - 500ms after the onset of the feature. In comparison of the p300 signals the greatest amplitude for each feature was compared regardless of latency within the observed window.

The feature importances or weights were observed with these three metrics, it follows that the variation from the metric for each subject could be compared to interpret some similarity between the measures. Through z-scoring each metric per individual, the relative weights could be observed (Figure 11). A z-score measure allowed the metrics to be compared as variance to each other. In this sense, a greater positive z-score corresponds to greater usage or weight of the feature and a more negative z-score corresponds to less usage. By comparing importance/weight values within metrics, it allows a more clear comparison of relative importance across metrics. Following z-scoring of each metric, the z-score values were aggregated by each metric across individuals and were compared via Spearman's rank correlation. A significant positive correlation was found between feature selection proportion and machine learning partial dependence, $r(43) = .71$, $p < .001$. A non-significant negative correlation was observed between feature selection proportion and maximum eeg amplitude, $r(43) = -.04$, $p = .814$. A non-significant negative correlation was observed between machine learning partial dependence and maximum eeg amplitude, $r(43) = -.07$, $p = .637$.

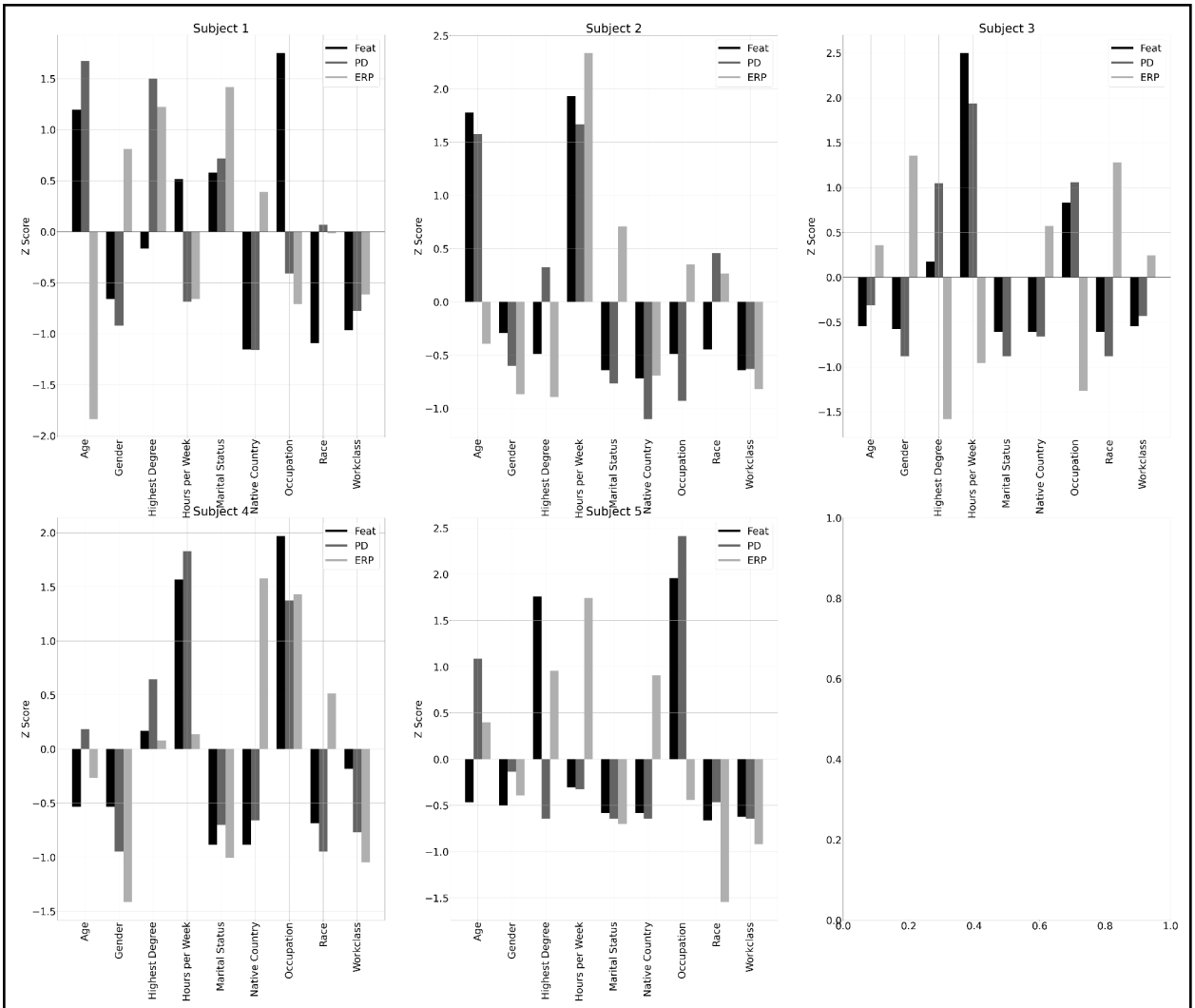

Figure 11: Bar charts comparing feature selection proportion, partial dependence, and p300 magnitude for each feature across all subjects. All measures were z-scored against themselves to compare all values as a measure of variation. Black bars indicate feature selection proportion, dark gray bars indicate partial dependence determined importance, and light gray bars indicate p300 determined importances.

Discussion

The results point toward interesting findings regarding human vs machine learning decision making. The performance of the human and the machine learning model on the same dataset was similar (difference of 0.06). However, as indicated by the TPR and TNR of both and ML, both have a slightly greater ability to discriminate between individuals that make more than $50,000 than individuals that make less than $50,000 relative to themselves. Despite this similarity, the percentage of unique correct answers points towards a non-overlapping similarity in performance (23% unique responses in subjects and 30% unique responses in the machine learning). This indicates that although there is similarity in the discrimination, there may still be differences arising from the approach that is not captured by the TPR and TNR. This possible

individual difference was further captured by the low percentage of problems in which subjects consistently were correct where machine learning was not correct; this indicates that there was not a significant set of problems that could be correctly handled by all individuals. This indicates that in a large group there are varying overlaps in expertise, but in a single individual to machine learning pairing, there tends to be a non-overlapping approach from either agent. To further build on single individual and ML performance, if the correct answers from both agents were combined, the resulting accuracy would be above the accuracy of either individual alone (18% - 24%). These observations illustrate that, in a difficult or incomplete set of problems/situations, a pairing of a human with a machine learning model could be explored to find an optimal solution, but the exact sets of problems covered by this pairing would vary by the individual paired with the machine learning.

The observation of machine learning in replicating the human's responses via the ROC curves shows some level of stability in supporting its feature weight distributions; if the machine learning can make sense of the human's response via its ability to predict the human's response, it leans towards some level of information processing from the human that presents itself as a pattern. Through observation of the model determined feature importances and the subject's explicit feature selection, it was apparent that individual's weigh information differently from each other and from the model. This pattern further supports the idea that there are individual differences in the problem solving approach, and in a sense, the machine learning is similar to another person or individual, albeit with a reliable pattern. It is necessary to draw attention to the subjects which machine learning was a poor predictor of, as is the case with Subject 5. The performance of the subject fell at 53% in the full feature task and 48% in the feature select task, this may indicate that the participants themselves were unable to find or had an unstable approach to the task; in this case, the model had difficulty replicating the performance as there was no clear approach for the model to take. This is reflected in the ROC curve of the unstable subject and the resulting importances determined by machine learning. The model produces a AUC close to 0.5 indicating near chance, in this instance, the resulting feature importances by the model are unreliable.

This feature weight was not observed in the p300 signals from participants. While there were differences within the magnitudes of p300s within individuals, there was very little similarity to the information the participant explicitly chose nor the features determined to be important by the machine learning model. A significant positive relationship was observed between the features selected by the individual and the features determined to be important to the subject by the machine learning model. Taken together it could be posited that machine learning is a relatively good predictor of how an individual weighs information explicitly, but the reliability of the neural response from the individual is inconclusive. In comparison with feature selection proportion and machine learning partial dependence, neither metric had a significant relationship with the p300 signal. It should be noted that in the scope of this paper, the signals were averaged across both sessions. It may be beneficial to understand the EEG signal better if

further segregation of the EEG signal took place, such as looking at beginning and end trials separately or separating and comparing EEG signals by values within feature categories.

Further data collected from more subjects may reveal clusters or subsets of feature weight approaches. As indicated by the machine learning partial dependence calculations, it is observed that some subjects have similar feature weights; subjects 1 and 2 as well as subjects 3 and 4 with both subgroups also having an overlap in hours per week (Figure 6). When comparing larger subject pools it may be possible and beneficial to recognize these subgroups. Future studies may also aim to further combine the differences observed between the machine learning model and human participants. Perhaps a third agent, a human or a separate machine learning model, may be able to take advantage of the differences in human and machine learning information weighting. In doing so, it may be possible to take a situation where both the human and machine learning operate poorly, and find an adjudicating system which maximizes performance above the performance of either agent individually.

References

Badman, R., Hills, T. T., & Akaishi, R. (2020). Navigating uncertain environments: Multiscale computation in Biological and artificial intelligence. https://doi.org/10.31234/osf.io/ced3t

Brown, S. (2021, April 21). *Machine Learning, explained*. MIT Sloan. Retrieved February 1, 2023, from https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2022, January 9). Retiring adult: New datasets for fair machine learning. arXiv.org. Retrieved February 1, 2023, from https://doi.org/10.48550/arXiv.2108.04884

Kirasirova, L., Bulanov, V., Ossadtchi, A., Kolsanov, A., Pyatin, V., & Lebedev, M. (2020). A p300 brain-computer interface with a reduced visual field. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.604629

Ma, N. (2017). Human Learning and Decision-Making, and Their Applications. *UC San Diego*. ProQuest ID: Ma_ucsd_0033D_16430. Merritt ID: ark:/13030/m5bw2b8x. Retrieved from https://escholarship.org/uc/item/18f6r7t3

Mansor, A. A., Mohd Isa, S., & Mohd Noor, S. S. (2021). P300 and decision-making in neuromarketing. *Neuroscience Research Notes*, *4*(3), 21–26. https://doi.org/10.31117/neuroscirn.v4i3.83

Mollick, E. (2022, December 14). *Chatgpt is a tipping point for AI*. Harvard Business Review. Retrieved February 1, 2023, from https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai

OpenAI. (2022, November 20). *CHATGPT: Optimizing language models for dialogue*. OpenAI. Retrieved February 1, 2023, from https://openai.com/blog/chatgpt/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in Artificial Intelligence. *Proceedings of the National Academy of Sciences*, *117*(48), 30033–30038. https://doi.org/10.1073/pnas.1907373117

Si, Y., Li, F., Duan, K., Tao, Q., Li, C., Cao, Z., Zhang, Y., Biswal, B., Li, P., Yao, D., & Xu, P. (2020). Predicting individual decision-making responses based on single-trial EEG. *NeuroImage*, *206*, 116333. https://doi.org/10.1016/j.neuroimage.2019.116333

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, *119*(11). https://doi.org/10.1073/pnas.2111547119

Trunk, A., Birkel, H., & Hartmann, E. (2020). On the current state of combining human and artificial intelligence for Strategic Organizational Decision making. *Business Research*, *13*(3), 875–919. https://doi.org/10.1007/s40685-020-00133-x

Wang, L., Zheng, J., Huang, S., & Sun, H. (2015). P300 and decision making under risk and ambiguity. *Computational Intelligence and Neuroscience*, *2015*, 1–7. https://doi.org/10.1155/2015/108417

West, D. M., & Allen, J. R. (2022, March 9). *How artificial intelligence is transforming the world*. Brookings. Retrieved February 1, 2023, from https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/

Won, K., Kwon, M., Ahn, M., & Jun, S. C. (2022). EEG dataset for RSVP and P300 speller brain-computer interfaces. *Scientific Data*, *9*(1). https://doi.org/10.1038/s41597-022-01509-w

Yun, Kyongsik & Chung, Dongil & Jeong, Jaeseung. (2008). Emotional Interactions in Human Decision Making using EEG Hyperscanning. International Conference of Cognitive Science. https://www.researchgate.net/publication/228466519_Emotional_Interactions_in_Human_Decision_Making_using_EEG_Hyperscanning

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). "An ideal human". *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1–25. https://doi.org/10.1145/3432945

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372852