CONFIRMATION BIAS ESTIMATION FROM ELECTROENCEPHALOGRAPHY WITH MACHINE LEARNING

THESIS

Micah Villarreal, Captain, USAF

AFIT-ENG-MS-19-M-065

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENG-MS-19-M-065

CONFIRMATION BIAS ESTIMATION FROM ELECTROENCEPHALOGRAPHY WITH MACHINE LEARNING

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Cyber Operations

Micah Villarreal, BS

Captain, USAF

March 2019

AFIT-ENY-MS-19-M-065

CONFIRMATION BIAS ESTIMATION FROM ELECTROENCEPHALOGRAPHY WITH MACHINE LEARNING

Micah Villarreal, BS

Captain, USAF

Committee Membership:

Dr. Brett J. Borghetti
Chair

Dr. Gregory J. Funke
Member

Dr. Michael E. Miller
Member

Abstract

Cognitive biases have been known to plague the human decision-making process for centuries. These biases often result in suboptimal outcomes in the face of uncertainty which can have disastrous effects in the fast-paced environments of military operators. Confirmation bias, which is the inappropriate bolstering of a hypothesis or belief whose truth is uncertain, can be especially harmful in military operations as information pertinent to alternative decisions is disregarded or downplayed with respect to information which supports the operator's current belief. Presently, there are two measures to estimate the degree of confirmation bias: 1) importance of information and 2) information selection. Unfortunately, these measures can be hindered by a multitude of subjective factors and cannot be collected fast enough to detect confirmation bias in real-time. This work investigates enhancing the current measures of estimating confirmation bias with behavior patterns and physiological variables.

In this pilot study, the MITRE-developed Assessment of Biases in Cognition (ABC) was completed by 15 participants. The ABC elicited biased behavior on decision making tasks while corresponding behavioral and physiological data was collected. To infer confirmation bias from brain activity, the relationship between electroencephalography (EEG) signals and behaviors associated with confirmation bias is modeled with machine learning. These models were utilized to classify the presence of confirming and disconfirming information. The artificial neural network achieved a classification balanced accuracy greater than 50% on two participants. However, overall model performance was low across all participants suggesting further research is

necessary. Although there was no significant difference in brain activity at the cross-participant level between the presence of confirming and disconfirming information, machine learning salient features in participants with relatively high machine learning performance were associated with brain locations that have been related to the presence of confirming information.

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Dr. Brett Borghetti, for his guidance and support throughout the course of this thesis effort. His insight and experience were certainly appreciated. I would also like to thank my committee members Dr. Funke and Dr. Miller for their support. Finally, I would like to thank my family, without their support and encouragement, none of this would have been possible.

Micah Villarreal

Table of Contents

List of Figures

List of Tables

CONFIRMATION BIAS ESTIMATION FROM ELECTROENCEPHALOGRAPHY WITH MACHINE LEARNING

I. Introduction

1.1 Motivation

In the face of uncertainty, the human decision-making process is known to suffer from cognitive biases which result in suboptimal outcomes [1]. The effects of these suboptimal decisions in military operations can be disastrous. In 1988, the commander of the USS Vincennes erroneously shot down a commercial Iran Air Flight resulting in the loss of 290 passengers' lives. The commander's error in judgement was partially contributed to cognitive bias which resulted in the commander relying too heavily on the wrong information [2].

Decisions made in military operations are particularly prone to cognitive biases due to the high stress, fast paced, highly uncertain environments military operators face on a daily basis. With the vast expansion of available information in the 21st century, the speed at which information is readily accessible has drastically increased, making the decision-making process more cognitively challenging than ever before [3]. With this ever-increasing volume of information readily available to military operators, the ability to detect suboptimal decisions from cognitive biases is necessary. More importantly, with the ability to objectively detect biased decisions, catastrophes resulting from poor decision-making processes can be prevented.

1.2 Problem Statement

Confirmation bias is the "inappropriate bolstering of hypotheses or beliefs whose truth is in question" [4] and is one of the most prevalent cognitive biases. This bias is crucial in military operations because it can result in overlooked information that is paramount to an optimal decision-making process. Current literature reports use of two measures to estimate the degree of confirmation bias: 1) information selection [5] and 2) importance of information [6], [7]. Unfortunately, these subjective measures can be hindered by a multitude of factors including evidence search strategies, evidence interpretation, socially acceptable outcomes, the participant's belief of what the experimenter wants to hear, and participant memory capacity [8].

This study intends to replicate the traditional confirmation bias measures of information selection and information importance while incorporating new behavioral and physiological measurements. Presently, there are no established methods to detect the presence of a confirmation bias in real-time as the established measures cannot be collected in real-time [9]. Mapping objective measurements, specifically behavioral and physiological, to the established behavioral measures will yield the slow and subjective measures unnecessary. Utilizing these mapped objective measurements, confirmation bias can be estimated as it is occurring, without the need for information selection and importance measurements.

1.3 Research Focus

This study will investigate decision-based confirmation bias relationships between behavior, self-reported information and psychophysiological signals collected when a

participant makes a decision while affected by confirmation bias. By modeling the relationships between behavioral data and physiology measurements from participants making decisions, this study will document the relationships exposing objective measures which might be used to identify confirmation bias.

1.4 Research Questions/Hypotheses

RQ1: During decision-making tasks, if the participant is required to make an initial decision, what impact does an initial decision have on participant behavior during subsequent information search?

> *Hypothesis: Making an initial decision before information search will result in bias which can be indicated by unbalanced information search behavior* [4]. *Similarly, if an initial decision is not made, there will be less bias and consequently a more balanced information search behavior.*

RQ2: What are the information acquisition behavior patterns associated with confirmation bias?

> *Hypothesis: Behavior patterns associated with a confirmation bias will be revealed by associating biased information selection* [10] *or information/question importance* [6] *with completion time, and information revisit..*

RQ3: Can a machine learning classification model using physiological signals estimate the presence of confirming and disconfirming information with performance greater than random chance?

Research Objective: Develop a machine-learning model able to classify the presence of confirming information with equal-class-weighted classification accuracy greater than 50%.

RQ4: Are neurophysiological signals in the right frontal lobe associated with confirming and disconfirming information? Are neurophysiological signals in the right frontal lobe salient features in a machine learning information classification model?

Hypothesis: In contrast to disconfirming information, confirming information will provoke increased activity in the brain's right frontal lobe [10] *which will be significantly different in neurophysiological signals. The difference in activity will result in features associated with the brain's right frontal lobe producing salient machine learning features.*

1.5 Methodology

Biased behavior is elicited in fifteen participants through decision tasks in a MITRE-developed Assessment of Biases in Cognition (ABC) platform. In each decision task, the participant selects confirming and/or disconfirming information (relative to the participants initial or final decision) to make a decision. A decision task is evaluated as biased, if the proportion of selected confirming information is greater than the proportion of selected disconfirming information. During the ABC assessment, behavioral and physiological measures are collected, including: information selection, completion time, information revisit, Electroencephalography (EEG), Electrooculography (EOG), and Electrocardiography (ECG).

The collected behavioral data is investigated to determine if any significant behaviors are associated with biased decisions and are suitable as machine learning features. For machine learning, the collected EEG data is segmented using two methods: by task and by information. In the task dataset, a machine learning classification model is trained to evaluate the decision the participant makes. The target variable is "biased" and the goal is to classify EEG signals as being from a biased or unbiased decision task based on how much of each type of information the participant selected. In the information dataset, a machine learning classification model is trained to evaluate each item of information the participant selects. The target variable is "confirm" and the goal is to classify EEG signal segments as either confirming or disconfirming, based on the information selected. In each dataset, the two types of features explored are a raw time series EEG signal, and in the frequency domain, the mean power of the five clinical frequency bands. All machine learning models are trained within-participant and un-tuned cross-validation metrics are reported due to the small number of observations per participant. Finally, model feature saliency is evaluated to determine important features for estimating confirmation bias.

1.6 Assumptions and Limitations

1.6.1 Assumptions

Given confirmation bias is dependent upon one's beliefs, there are some key assumptions that must be made to adequately assess the degree of confirmation bias and label information as confirming or disconfirming.

- Beliefs held prior to the ABC assessment will not affect the participant's decision in the decision-making tasks. If a prior belief is held, the belief will be reflected in the initial decision.
- The initial decision made by the participant establishes a belief in the participant prior to information search. The established belief is suitable for labeling information as "confirm" or "disconfirm" during information search. In decision tasks without an initial decision, the participant's belief is not known prior to information search. In these tasks, the final decision made after viewing information is suitable for labeling information as "confirm" or "disconfirm" during information search.
- Each participant will have biased and unbiased decision tasks in the ABC assessment.
- EEG activity is different in the presence of hypothesis-confirming and hypothesis-disconfirming information.
- The participant is not aware of the ABC assessment content and does not have prior knowledge of the nature of the experiment.
- The participant will complete the ABC assessment to the best of their ability.

1.6.2 Limitations

The participant demographics for the experiment were exclusively volunteers from the Air Force Institute of Technology. All participants were male, United States military or federal government civilian personnel, with at least a Bachelor's degree. The mean age was 29.4 years with a standard deviation of 7.28 years. The lack of diversity in the

participant pool indicates the results in this work may not be generalizable outside of this demographic.

The decision tasks employed in the ABC assessment are complex decisions that require reading and processing large amounts of information. To prevent participant fatigue, the ABC assessment only consisted of fourteen decision tasks. With such a limited number of tasks, few observations were available from each participant. Additionally, there were not enough decision tasks of the same type such that a training, validation and test set could each include biased and unbiased decision tasks. Due to the small volume of observations for machine learning, a test set was not utilized for reporting machine learning performance metrics as is traditionally desired. All reported machine learning performance metrics are cross-validation metrics. To prevent inflation of the cross-validation metrics, this work employed only simple, un-tuned, machine learning implementations. With a larger dataset, machine learning models could be tuned and better-performing models could be developed and evaluated using a sequestered test set.

1.7 Contributions

This work contributes to decision-making research by augmenting the traditional measures employed to assess a biased decision with behavioral and physiological measurements. At the time of this work, estimating confirmation bias with machine learning has never been explored. Multiple facets of estimating confirmation bias through behavioral and physiological measurements are investigated, building a foundation for future work to build on.

The exploratory machine learning approach applied in this work indicates the best method for identifying the presence of confirmation bias from EEG data is to classify the position of information as confirming or disconfirming. On the information dataset with frequency features, the artificial neural network obtained above 50% baseline balanced accuracy on 2 of the 15 participants with the highest balanced accuracy on a participant being 62.6%. Although these early results are not reliable enough for operational use, they suggest there may be a relationship between EEG signals and the presence of confirming information. In addition, features from the brain area associated with the presence of confirming information were salient features in the highest performing participant models. The F4, F6, and F8 features associated with the right frontal lobe of the brain were one of the eight most salient features (out of 320 features) in four of the participants with the highest random forest classifier performance. On the task dataset with frequency features, no participants had a model obtain above the 50% baseline balanced accuracy.

Although machine learning results on the information selection with frequency features dataset are only marginally better than the chance, they do suggest it is possible to classify the presence of confirming and disconfirming information from EEG signals. By classifying the position of information selected in a decision, it is possible to estimate a decision with confirmation bias from physiology signals by detecting when more confirming information is selected during decision making.

1.8 Preview

This work consists of six chapters. Chapter II reviews present literature on confirmation bias and decision making. Specifically, it focuses on the traditional methods applied to detect a biased decision. Lastly, this chapter examines common machine learning methods utilized in EEG classification. Chapter III outlines the methodology for the human-subject experiment implemented to collect behavioral and physiological data as well as the implemented machine learning pipeline. Chapter IV discusses the analysis and results of the behavioral data and the machine learning performance metrics on the physiology data. Lastly, Chapter V concludes this work by answering the research questions with results and providing recommendations for future confirmation bias estimation research.

II. Literature Review

2.1 Chapter Overview

This chapter provides a summary of decision-making research on confirmation bias. Confirmation bias definitions, measures, and task environments used in research are discussed. In addition, a brief overview of applicable machine learning methods is provided.

2.2 Definitions, Themes, and Concepts

The grave impact of military decision-making can be illustrated in the early roots of American history. General Robert E. Lee, the commander of the Confederate forces was a highly successful leader, but in the battle of Gettysburg he was defeated because he underestimated his opponent. He was said to believe that victory would come due to his own doings. Many historians attribute Lee's poor decision-making process in overvaluing information supporting his belief of victory due to confirmation bias [11]. A short time after Gettysburg, George Armstrong Custer's decisions at the Battle of Little Bighorn in 1876 resulted in significant loss of life with over 200 troops lost. Custer had significant success in the Civil War and was over-confident in his decisions against an alliance of Plains Indians at the Battle of Little Bighorn. Despite his plans breaking down during the battle, Custer anchored to his original plan and failed to reassess the situation [12]. Although cognitive biases have been known to result in suboptimal decisions for centuries, they are still prevalent in decision-making today.

Today with an abundance of real-time information available to military operators, the recognition and prevention of error due to cognitive biases is even more important as

today's military operators have substantial information resources at their disposal such that the failure to rely on this information appropriately, rather than the lack of information, is more likely to result in a poor decision-making process that results in an increased probability of an undesirable outcome. In 1988 the commander of the USS Vincennes received conflicting information on the type of an approaching aircraft and shot down Iran Air Flight 655 killing 290 passengers. The commander erroneously believed the aircraft to be an F-14 fighter from the Iranian Air Force. One cause of the error in decision is believed to be the high tension and recent incidents which caused the commander to suffer from confirmation bias and overvalue the information which supported the hypothesis that the aircraft was a hostile military airplane [2]. While this is one of the most prominent examples which have led to substantial loss of life, all aspects of military operations entail decision making. From pilots constantly assessing their rapidly changing state to make quick decisions, to intelligence analysts assessing pertinent information, to cyber analysts correctly identifying a cyber-attack, all situations require objective assessment of information to make timely, unbiased decisions.

When making decisions under uncertainty, people often use heuristics, or mental shortcuts, to navigate and simplify complex decisions [1], [13]. For example, objects that are closer in distance appear more clear than objects at a far distance; consequently when objects are clear we often over-estimate how close the object is [1]. These innate strategies to use heuristics are effective part of the time, but also result in consistent errors [14]. This phenomenon, or unconscious error in judgement, which results in a suboptimal decision-making process is a cognitive bias. Cognitive biases are not only prevalent in arbitrary contexts, like the previous mentioned distance estimation example,

but are widespread in real-world contexts including, but not limited to, national policy, intelligence analysis, medical practices, the judicial process, and science [4].

There are numerous cognitive biases which can significantly impact the decision-making process. Some of the most prominent cognitive biases include anchoring, availability bias, and confirmation bias. The availability bias occurs when people over-estimate the probability of an event because of their ability to easily recollect an instance judged similar to the event from memory [1]. For example, people may overestimate the probability of winning the lottery because they can easily recall a memory of recent lottery winners reported on the news.

During the pre-decision stage of the decision-making process, people tend to make estimates based on initial values. Anchoring bias occurs when people anchor on these initial estimates and fail to properly adjust these estimates in light of new information prior to making their final decision [1]. The canonical example of anchoring occurs when buying a used car: whoever makes an offer first will set the initial value. If the seller offers to sell the car for $8,000 the buyer may experience an anchoring bias if they anchor on this initial value and offer close to $8,000 despite their previous valuation of the car.

Lastly, confirmation bias is the inappropriate bolstering of a believed hypothesis in the face of uncertainty [4]. In a police investigation, confirmation bias occurs if the investigating officer forms an initial hypothesis on who they believe the guilty suspect is and consequently only searches for evidence or overvalues evidence which supports their hypothesis. Confirmation bias can be especially damaging to analysts because pre-

conceived beliefs can result in missed or misinterpreted information. For this reason, this work explicitly focuses on confirmation bias.

Given cognitive biases can have a significant impact on the decision-making process, clearly expanding the decision-making process can help with understanding cognitive biases. The decision-making process can be split into three stages which are:

1) Pre-decision
2) Point of decision
3) Post-decision

The pre-decision stage is the point of basic information gathering where conditions are proposed and alternatives are generated. The point of decision is when one of the previous alternatives is chosen and commitment to the decision is made. Lastly, the post-decision stage is when rationalization of the decision occurs and seeking of more information may be biased toward the previous decision [15]. Cognitive biases can occur at several different stages or throughout all of the stages of the decision making process, or even between multiple decisions [16]. For example, memory biases, which are biases that affect the process of recalling information, affect the pre-decision stage of the decision-making process. Whereas cognitive dissonance, which occurs when one has inconsistent thoughts about their decision and as a result believes an alternative decision was better, occurs in the post decision stage [13].

The confirmation bias generally affects the first two stages of a decision. When a belief is held, confirmation bias is present during the pre-decision stage if basic information gathering is biased towards the held belief. The biased search or overvalue of belief confirming information then impacts the point of decision. In literature,

participants are generally primed to hold a belief by completing one iteration of the decision-making process; confirmation bias is then measured in a subsequent decision with respect to the first decision.

Throughout the body of this work, the terms “confirming” and “disconfirming” are utilized to express the relationship between the information and a participant’s current belief or hypothesis. Confirming information is information that confirms or supports the participant’s current belief. Disconfirming information is information that contradicts or disconfirms the participant’s current belief. Neutral or irrelevant information is information that neither confirms nor disconfirms the participant’s belief. The basis of how one interacts and assigns value to these types of information is the crux of quantifying confirmation bias.

2.3 Confirmation Bias

2.3.1 History and Competing Definitions

The prevalence of the confirmation bias is illustrated by one of the first documented accounts being over 400 hundred years ago when Sir Bacon expressed “The human understanding, when any proposition has been once laid down . . . forces everything else to add fresh support and confirmation…” [17]. Although the tendency for people to exhibit these behaviors have been known for decades, some of the most notable research began circa 1960 with the Wason abstract rule discovery experiment which exhibited error in hypotheses by seeking confirmatory evidence [18]. In his work, Wason gives participants a sequence of three numbers and states the three numbers follow some unspecified rule. The goal of the participants is to determine this rule by choosing

minimal sets of three numbers. After each set, participants are informed if the set conforms to the rule. At any point, participants are allowed to declare what they believe the rule to be. For example, suppose the governing rule the participant was supposed to determine was "sequences which contain only integers increasing in order of magnitude". Then the experiment administrator provided the sequence "2, 4, 6" as a sequence which is compliant with the rule. Next the administrator asks the participant to provide other sequences for validation, with the ultimate goal of determining what the rule is. A participant may choose the number sequence "8, 10, 12" and the administrator would inform the participant the numbers conform to the rule. The participant may then choose the number sequence "3, 2, 1" and would be informed the numbers do not conform to the rule. A plausible, but incorrect rule, the participant may declare at this point is "sequences of even numbers in increasing order". The participant continues this process of choosing sets of numbers until the correct rule is declared. Wason found participants who arrived at the correct rule on their first rule guess did so by testing many confirming and disconfirming number instances before guessing. While those who announced an incorrect rule in their first guess, did so by testing only a small number of confirmatory instances. The significance of Wason's work is people often create erroneous hypotheses when only seeking confirmatory evidence.

There are many competing definitions in confirmation bias research. This ambivalence in what true confirmation bias is, has led to many misconceptions and disagreements on empirical findings. The positive test strategy, as proposed by Klayman and Ha, suggests that many of the empirical findings that are classified as a "confirmation bias" align more closely with a positive test strategy rather than a confirmation bias. The

positive test strategy is articulated as: when testing a hypothesis, the tendency for people to seek cases that are believed to demonstrate the event rather than conditions that are thought to lack the event. This strategy is believed to be a good heuristic for testing the truth of a hypothesis despite the fact it can lead to consistent errors [19]. One important characteristic of this perspective, is that despite the utility of the positive test strategy, it is still prone to errors and can ultimately lead to confirmation bias.

Another framework that analyzes confirmation bias is the Bayesian perspective proposed by Fischhoff and Beyth-Marom. The Bayesian perspective utilizes an inferential approach using Bayes theorem to show how the framework can be applied to the evaluation of hypotheses. By applying the Bayesian perspective to empirical findings on confirmation biases, Fischhoff and Beyth-Marom declare peoples' intuitive inferences that lead to deviations from the Bayesian model are better classified with this framework rather than a confirmation bias.

In a study on preferential search for hypothesis-confirming behavior, Snyder and Swanson conduct an experiment in which their participants are instructed to select questions to ask another person to test either the hypothesis that the person is an extrovert or the hypothesis that the person is an introvert. By participants selecting a majority of questions that would confirm the assigned hypothesis (extrovert or introvert), Snyder and Swanson conclude that people tend to have a preference for a hypothesis-confirming strategy in testing hypotheses [20]. Fischhoff and Beyth-Marom emphasize that, from a Bayesian perspective, there is no possible way to ask questions such that all of the possible answers would be supportive of a particular hypothesis. Therefore, asking questions that are seemingly supportive of a particular hypothesis is not in fact a

confirmation bias. From the Bayesian perspective, Snyder and Swans findings from asking questions that would likely confirm the hypothesis, are non-diagnostic questions; thus, choosing a non-diagnostic hypothesis confirming question is not confirmation bias.

One of the most notable works on the confirmation bias is Nickerson's research in which he establishes what he believes to be a confirmation bias and provides a working definition for present research [9],[13], [21]. Nickerson takes the confirmation bias to be a generic concept that entails several other notions that "connote the inappropriate bolstering of hypotheses or beliefs whose truth is in question" [4]. Confirmation bias can be motivated or unmotivated. An example of a motivated confirmation bias is when someone rates evidence on the effectiveness of the death penalty (confirming evidence) as more convincing because their opinion is the death penalty is effective, whereas an unmotivated confirmation bias would be if the participant had no opinion on the death penalty, yet rated one type of evidence as more convincing when both are equally supportive [22].

Through examination of the empirical findings at the time of Nickerson's work, he outlines five main findings in regards to information search. When a hypothesis is favored people tend to:

1) Restrict attention to the favored hypothesis
2) Give preferential treatment to evidence supporting existing beliefs
3) Look for primarily positive cases that support the hypothesis
4) Overvalue these positive confirmatory cases
5) See what they are looking for

First, people tend to restrict attention to a favored hypothesis. This narrowed focus will often cause the interpretation of data to be one that supports the favored hypothesis and leads to a failure to recognize that the evidence supports an alternative hypothesis more than the favored hypothesis.

The second finding is similar to restricted attention and is the preferential treatment of evidence supporting existing beliefs. Preferential treatment of evidence materializes when one gives more importance to evidence that supports their belief than evidence that contradicts the belief. One may not completely ignore the unsupportive evidence but are less receptive and may explain away the evidence.

Third, people tend to look for primarily positive cases, or conditions in which their hypothesis would be supported despite the hypothesis being related to a vested interest or not. These positive cases may not always be logically confirmatory, but rather psychologically confirmatory. Psychologically the person believes the case will confirm their hypothesis when in fact logical confirmatory evidence would be testing the hypothesis to be correct through confirming and disconfirming evidence.

The fourth finding is overvaluing positive confirmatory instances of a hypothesis. This is very similar to the second finding but is different in the sense that overvaluing positive confirmatory instances results in the acceptance of a hypothesis with less confirming evidence than the rejection of the hypothesis with inconsistent evidence. For example, one is more likely to accept a hypothesis they favor with only two supporting pieces of evidence than to reject the hypothesis with four contradicting pieces of evidence.

The last of Nickerson's general findings is people tend to see what they are looking for. One example of this sensation is illustrated in a study in which two groups watched a video of a child taking a test. In the study, one group was led to believe the child had a low socioeconomic background while the other group was led to believe the child was of high socioeconomic background. The group who believed the child had a high socioeconomic background rated the child has a high performing student while the other group rated the child as a poor performing student [4]. These effects can greatly impact the decision-making process and induce confirmation bias if an initial hypothesis is formed. This work utilizes Nickerson's definition of confirmation bias along with his five findings as a working definition of confirmation bias.

2.3.2 Theory of Confirmation Bias

Many of theories behind the bias can be separated into two main categories: 1) task environment theory and 2) cognitive process theory [23]. The task environment theory encompasses many of the motivated heuristics, like the positive test strategy. This theoretical approach is formed on the basis that biases occur because a heuristic is misapplied and results in an error. The heuristics are developed as adaptations to environments as a means of efficiency or survival, not because of one's inability to process large amounts of information. For example, using the positive test strategy, if the task environment rarely yields a contradicting instance of the hypothesis, it is generally more efficient to test a confirming instance of the phenomenon. The positive test strategy serves well but can result in biased behavior when misapplied to an unfamiliar task environment. Despite the presence of the task environment theory, the majority of theories in literature fall under the cognitive process category.

The cognitive processes category encompasses numerous different theories ranging from cognitive limitations to cognitive dissonance. Evans postulates the confirmation bias stems not from a motivation to seek confirming evidence but rather people are unable to think of a way to falsify due to cognitive failure [24]. A similar theory is the structure and human thought fosters confirmatory strategies because it is easier to think of ways that confirm a hypothesis than the contrary [20]. Tversky and Kahneman's eminent work on cognitive biases also suggests biases occur because cognitive limitations necessitate the use of heuristics in the decision process. In addition, people tend to look for similar features rather than distinctive features. In decision making, where it's easier to think of a confirmatory instance of the hypothesis, the similar evidence is evidence that confirms the hypothesis [1], [25]. Another theory, which falls under the cognitive processes category, is the thought that confirmation bias occurs to decrease cognitive dissonance [23]. By searching for evidence that confirms the initial hypothesis, there will be less dissonance post decision because confirming evidence supports the final decision.

While the discussed cognitive theories are by no means exhaustive, the abundance of theories illustrates the degree to which the underlying causes of cognitive biases are unknown. The ambiguity in the underlying mechanism of a confirmation bias indicates research with new measures to detect the presence of a bias may aid in determining the true underlying mechanism which results in biases.

2.3.3 Measuring Confirmation Bias in Decision Making

The standard approach to measuring confirmation bias in decision making is a three step process [5], [10], [26]. In this approach participants:

1) Perform a task and make an initial hypothesis.
2) Review additional evidence or information.
3) Make a final hypothesis.

In the initial hypothesis stage of the experiment participants are given a summary with incomplete evidence. After reviewing the initial evidence, the participant is asked to form an initial hypothesis by way of pre-determined choices or other means. The formation of the initial hypothesis allows for a baseline to measure the degree of bias present in the ensuing information review.

After making an initial decision, the participant is presented with some new or additional information that will have an impact on the participants working hypothesis. The new evidence is either neutral, confirming or disconfirming information with respect to the participant's initial hypothesis. The presence of all types of evidence allows the experiment to determine if the participant's examination of the information is neutral or biased. Upon completion of the information review, a final decision is made which allows the effect of the evidence and initial decision on the final decision to be measured.

2.3.3.1 Information Search

One measure of confirmation bias is selective information search [7]. When experiencing confirmation bias in information search, people selectively search for confirming information. The degree of confirmation bias is quantified by the difference of confirming and disconfirming information reviewed or selected. Information search has been implemented as a means to measure confirmation bias in multiple decision making tasks including: financial decision tasks, health policies, and intelligence analysis [5], [27], [28].

Mynatt et al. demonstrate selective information search by having participants attempt to discover laws governing the motion of the particle in a computer simulation. Initially participants interact with one environment which is oriented so the majority of participant's favor one initial hypotheses. Upon making a hypothesis, participants choose five more environments to fire particles in. The selective information search was demonstrated by the fact that 70% of the participants chose to conduct further tests in environments that would confirm their initial hypothesis [26].

2.3.3.2 Information Importance

Another prominent measure of confirmation bias is the importance, or value, of confirming evidence [6], [7]. The degree of confirmation bias experienced manifests through information importance by giving higher value to confirming information than disconfirming information [7]. Information importance is quantified through two different methods by prompting participants to assess: 1) what information is most important to answer a given question or by 2) what information was most important in making a final decision.

In assessing information importance regarding a question, participants are provided with a question or hypothesis to investigate. A collection of confirming and disconfirming statements is provided. The participant then must choose what information, if investigated further, would be most beneficial to answering the question. Results indicate people experiencing bias systemically rate confirming information as most important [23].

Consistent with Oswald and Nickerson's view on the confirmation bias, numerous studies use the importance of confirming and disconfirming information as a means to

evaluate the presence of a bias. Lehner et al. conducted a study in which they measure confirmation bias by importance of evidence in a complex intelligence analysis task. In this study, participants exhibited a positive correlation between a preferred hypothesis and assessment of confirming evidence. More succinctly, when a hypothesis is favored, confirming evidence was assessed to be more important than disconfirming evidence [6]. These results indicate greater assessed information importance for confirming information than disconfirming information may be used to measure the degree of confirmation bias present.

2.3.3.3 Physiological Measurements

Electrophysiological measurements are objective external measurements including electroencephalography (EEG), electrocardiography (ECG) and electrodermal activity (EDA). These measurements can provide insight into neurological and psychological activity which is beneficial in understanding confirmation bias. One particular work which is unique in measuring confirmation bias is Minas et al.'s work which employs electrophysiological signals in addition to the traditional measures [10]. This work detects differences in electroencephalography (EEG) and electrodermal activity (EDA) associated with the presence of confirming information. Although electrophysiological measurements associated with confirmation bias is still in its infancy, these results are the first objective measurements of confirmation bias.

2.3.3.3.1 Electroencephalography

Electroencephalography (EEG) is the physiological measure of electrical activity on the scalp which captures electrical activity from the brain. EEG can be more useful than behavioral measures because of its high degree of sensitivity which allows

distinction of cognitive processes [29]. Activation of the right frontal cluster of the brain through EEG has been correlated with the presence of hypothesis confirming information compared to disconfirming and irrelevant information [10]. These results are significant because presently the only viable methods for measuring the presence of confirmation bias is through behavioral measures; subjective information importance and information selection. The correlation of right frontal cluster activity through EEG provides an objective way to detect confirmation bias and could be used to detect a bias in near-real time. Identifying the presence of bias from EEG signals would allow a bias to be detected without the current behavioral measures.

2.3.3.3.2 Electrodermal Activity

Electrodermal activity (EDA) is the measure of changes in electrical activity on the surface of the skin. EDA measures the sweat response and is a common index into the implicit emotional states as the measured psychophysiology variable is not contaminated by explicit activity [30]. The Galvanic Skin Response (GSR) is typically measured on the hand and this signal increases with increased sweat gland activity. GSR is a useful measure in decision making because it can reflect the anticipation of negative outcomes or unconscious emotional changes during the decision-making process [31]. Relative to arousal in the presence of disconfirming information, the presence of hypothesis-confirming information has been associated with increased arousal measured six seconds after confirming information onset [10]. These results indicate GSR during a decision-making process can provide an objective measure of emotional response to the presence of confirming or disconfirming information.

2.4 Machine Learning

Machine learning is the science of programming computers to learn from data [32]. The canonical example of applied machine learning is a spam filter which flags emails as spam. The spam filter is a machine learning model which takes emails as an input and outputs if the email is spam. For the model to learn, training and validation data sets are used. A training set in this instance is a set of emails with a label for each email as spam or not spam. The training set is the data from which the model learns patterns which are associated with specific labels. For example, the model may learn to associate the words “free” and “money” in emails with the label “spam”. After training the machine learning model on the training data, the model performance is tested with unseen data or validation data. The model’s performance on the validation data can be used to tune parameters in the model to improve performance. Finally, the machine learning model’s true performance is tested on test data which has not been seen by the program or exploited for tuning parameters. Applying the model to the test data provides a method to assess expected model performance in real operation.

Broadly speaking, machine learning problems can be categorized as supervised or unsupervised. In supervised learning, a label is available, whereas in unsupervised learning there is no label available. A label is the category an observation belongs to and is the output of the model. In the spam email example, every email has a label of “spam” or “not spam” and is supervised learning because the label of the email is known. In addition, problems solved with machine learning can be classification or regression problems. Generally, in a regression problem the goal is to predict a number or the output of the program is numerical in nature. For example, predicting the price of a house based

on features like location, size, and number of bedrooms. A classification problem is one with the goal of predicting a categorical output. The spam email example is a classification problem because the objective is to classify emails as spam or not spam. Using the previous definitions, the spam email detection problem is a supervised classification machine learning problem. The remaining portion of this section covers the different machine learning algorithms that are utilized in this work.

2.4.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a linear method for classification which uses a linear decision boundary to differentiate between classes [33]. LDA has several advantages over other linear models which are: it is stable when classes are well-separated, it is stable with a small number of observations, and is suitable for more than two classes. LDA approximates a Bayes classifier by operating on the assumptions that each class has an approximate Gaussian distribution with a class-specific mean vector and a common covariance matrix. These assumptions yield the discriminate function below [33]:

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}^2{}_k}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

$\hat{\mu}_k$: mean of class k

$\hat{\sigma}^2$: weighted average of sample variances for each K classes

$\hat{\pi}_k$: proportion of training observations that belong to class k

Simply stated, LDA determines the probability that observation *x* belongs to each class and assigns *x* the class with the highest probability based on the discriminate function.

Although LDA is not the most commonly applied machine learning model in EEG classification, it has been shown to perform reasonably well. In the work by Binias et al., LDA was utilized to discriminate between EEG signals from aircraft pilots brain activity pre-event and post-event. In the pre-event class, pilots were focused and anticipating presentation of a visual cue whereas in the post-event class pilots were reacting to the visual cue. In classifying the pre and post-event states of brain activity of aircraft pilots from EEG signals, LDA performed the second best among all models with a mean accuracy of 73.01% for the two-class problem [34]. The LDA model for this application outperformed support vector machines, random forest, and k-nearest neighbor's models but fell short of the artificial neural network model performance. These results indicate LDA should be considered in EEG classification.

2.4.2 Random Forests

Random forests classifier is an ensemble machine learning method which uses decision trees in its ensemble [32]. An ensemble method is a classifier which combines outputs from multiple algorithms to classify the output. Random forests are different from the previously discussed machine learning methods in that decision trees are the basis of the model which can classify linear and non-linear classes. Decision trees split the classes into subgroups at discrete points of distinguishing features. The random forest ensemble of decision trees is created by making a set number of decision trees with random different subsets of the available features in each tree. The use of random subsets of features in each tree de-correlates the trees and allows the overall prediction to be more reliable [33]. For any input, the random forest method averages the prediction of

each of the decision trees and the class with the majority of votes is the resulting class of the input.

Although random forest classifiers usually do not yield as high performance as neural network methods in EEG classification [35], they allow salient features to be identified. Random forest classification methods can provide further insight into features which allow for distinction between classes. Random forest classifiers have obtained an accuracy of 75% in a EEG binary classification problem for which the classes in questions were brain activity under concentration and brain activity during meditation [36]. While this performance is not state of the art for EEG classification, the accuracy is above random chance and can provide insight into salient features.

2.4.3 Artificial Neural Networks

In general, neural networks are formed by stacking layers of nodes on top of each other which allow the overall network to model non-linear relationships. These layers are connected to each other and have weights which are modified during the learning process. The weights are changed until the network correctly maps an input to the desired output. Initially, weights are randomly initialized and the network performs poorly. To measure how far the networks current mapping is from the desired output, a loss function is used. An optimizer then takes the measure provided by the loss function and implements the backpropagation algorithm to adjust the layer weights [37]. Through many iterations of this process, weights are adjusted until inputs are mapped to the desired output.

The simplest neural network is the fully-connected neural network or artificial neural network (ANN). In ANNs, every node in a given layer is connected to each node

in the subsequent layer, which results in slow training time for large ANNs. The two other common types of neural networks this work utilizes are convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

2.4.4 Convolutional Neural Networks

Convolutional Neural Networks (Convnets) are a class of neural networks which learn local spatial patterns in the input data. These local patterns are learned by convolving filters with specific patterns over the input data which outputs encoded aspects of the input data known as feature maps. By stacking layers, the network can learn hierarchical patterns from the input layer [37]. To down-sample the outputs from convolutional layers and learn larger spatial features, a max-pooling layer conventionally follows a convolutional layer with a max tensor operation which halves the feature maps. One of the main advantages of Convnets is the weights of connected layers are shared across a given layer which reduces the overall parameters of the model and consequently training time and overfitting.

Convnets are known for state-of-the-art performance in image classification because of their ability to recognize local patterns. These characteristics are frequently exploited for classification of brain activity from EEG signals. In classification of motor movements in the right hand or right foot from EEG signals for brain-computer interfaces, Shang et al. create sparse representations of EEG features and translate these features to a two-dimensional signal which is input into a Convnet. With this implementation of Convnets in EEG classification, an average accuracy of over 80% was obtained for the two-class motor movement recognition [38].

2.4.4.1 Temporal Convolutional Networks

Temporal Convolutional Networks (TCNs) are a specific CNN architecture that has been shown to have a longer effective memory and faster training times than RNNs in sequence modeling [39]. The general structure of the TCN architecture inspired by Bai et al. is shown in Figure 1 [39]. A dilated convolution is a convolutional layer with a dilated kernel. The standard convolutional layer has a dilation of 1 as illustrated by the first hidden layer in Figure 1(a). The second layer in Figure 1(a) has a convolutional layer with a kernel dilation of 2. This dilated kernel has a coarser input from the previous layer as the kernel is convolved over alternating nodes from the previous layer. The benefit of using dilated convolutions is that when dilations increase exponentially and are stacked, the receptive field of the model increases exponentially while the number of parameters increase linearly. A residual block consists of a dilated Conv1D layer with 'causal' padding, a Rectified Linear Unit (ReLU) activation layer, channel normalization, spatial dropout of 0.05 and a Conv1D layer with 'same' padding as shown in Figure 1 (b). Channel normalization is completed using the max activation of the ReLU activation layer. A skip connection from the input to the residual block is element-wise added to the output of residual block. This residual connection allows the linear sequential relationships from the input to be maintained while the Conv1D layers learn non-linear patterns in the data. This architecture prevents the relationship from the current time step and earlier time steps in the data from being lost due to a deep architecture.

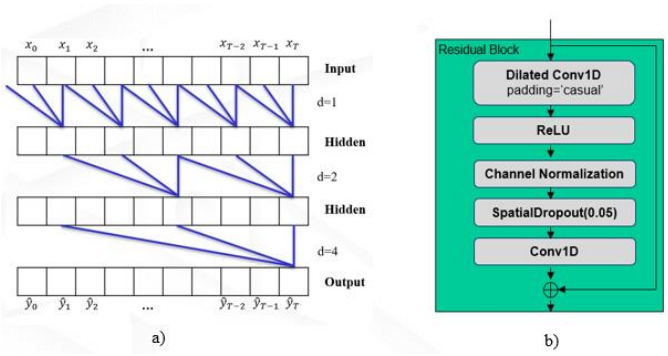

Figure 1: TCN architecture (a) Dilated causal convolutions (b) Residual block

2.4.5 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a form of neural networks which maintain a state or memory which allows the network to learn sequences. The internal state is maintained by a recurrent connection to itself which enables the network to process the current element of an input based on the element as well as the previously seen elements [37]. A major problem with traditional RNNs is they suffer from vanishing gradients. The vanishing gradient problem is without explicit memory blocks, traditional RNNs are unable to retain information from distant time steps in large sequences. There are two prominent types of RNNs which do not suffer from vanishing gradients: Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). In short, LSTMs and GRUs allow important information from distant time steps to be used in the current time step predictions and irrelevant information to be forgotten [37].

RNNs are prominently used for machine learning problems where predictions are sensitive to time order. The temporal aspects of EEG make classifying brain activity from EEG signals a natural application of RNNs. RNNs have been shown to have significant improvement over other machine learning methods in classifying vastly different types of

brain signals from EEG signals. In classification of six hand motions from a grasp-and-lift experiment, the RNN obtained a correct classification of 94.8% across all classes which reduced the test error by 23.5% compared to non-RNN networks [40]. In addition, RNNs have also been shown to obtain the lowest test error in mental load classification from EEG signals. Specifically, the addition of LSTM layers in a Convnet reduced the test error for a four class mental load classification by 21.5% [35].

Even more notable, LSTMs have been used in architectures to classify cross-participant EEG brain activity of high and low workloads. Hefron et al. created a multi-path convolutional recurrent neural network (MPCRNN) to classify EEG of brain activity of high and low workload with up to 86.8% accuracy on EEG from participant data which was unseen by the model [41]. The implemented MPRCRNN architecture consisted of CNNs and LSTMs, which outperformed LSTMs and CNNs by themselves. The application of RNNs in EEG signal classification has been shown to be versatile for varying types of brain activity classification; attaining state of the art performance. These results suggest RNNs may perform well in a biased classification task from EEG signals.

2.5 Conclusion

The confirmation bias is a prominent cognitive bias which results in systemic errors in decision-making due to the inappropriate bolstering of a believed hypothesis. Prior research used subjective measures to quantify confirmation bias which include undue information importance and information selection. These subjective methods can be unreliable and cannot be assessed during the decision process for real-time detection without disruption. Subjective measures are inherently unreliable and slow - meaning

real-time detection of a bias in unachievable. To objectively measure a bias, the selection of information with undue importance during the decision process should be quantified without self-reporting feedback. One possible avenue to attain an objective measure is through physiology signals. If specific signals can be associated with the presence or absence of confirmation bias, a bias can be detected without the necessity of self-reporting assessments. Currently, there is little research on using behaviors to measure confirmation bias, but EEG and EDA signals have been correlated to the presence of preference supporting information. These findings indicate there may be credence to measuring a bias through physiological signals. Neural networks are the current state-of-the-art for performance in classifying brain activity in EEG signals. This indicates neural networks may perform well in classifying the presence of a bias from EEG signals if there is specific brain activity associated with confirmation bias. Further research on associated physiological signals is necessary to explore the feasibility of measuring and detecting the presence of confirmation bias objectively.

III. Methodology

3.1 Chapter Overview

The objectives of this chapter are to outline a human-subject decision-making experiment and describe a machine learning approach to detect confirmation bias. The results of applying machine learning to the collected data will aid human machine teaming research in the feasibility of real-time detection of biased decisions. If biased decisions can be detected, human machine teaming agents may be able to aid operators by mitigating biased decisions.

The next section of this chapter establishes the research questions and proposed hypotheses, which will be investigated. The third section outlines the decision-making experiment methodology using the Assessment of Bias in Cognition (ABC) test. The experiment methodology establishes the independent and response variables, participant demographics, materials, procedures and the analysis strategy. The fourth section of this chapter encompasses the machine learning approach that will be applied to the data collected from the experiment. Finally, this chapter concludes with a summary of the covered topics.

3.2 Research Questions

Present research on confirmation bias measures a bias in the decision-making process by priming the participant prior to the decision. This priming is either implemented through an initial decision prior to the measured decision or by providing an accepted hypothesis [20], [23]. Since people tend to look for primarily positive cases

which support a hypothesis, whether or not they have a vested interest [4], the initial priming results in confirmation bias. To model the relationships between behavioral measures and physiological measurements balanced data is essential for high performance models. Consequently, collecting data on biased and unbiased decisions is necessary for robust models. This yields the following question and hypothesis:

> **Research Question 1**: During decision-making tasks, if the participant is required to make an initial decision, what impact does an initial decision have on participant behavior during subsequent information search?
>
> *Hypothesis: Making an initial decision before information search will result in bias which can be indicated by unbalanced information search behavior* [4]. *Similarly, if an initial decision is not made, there will be less bias and consequently a more balanced information search behavior.*

To assess research question one, the Assessment of Biases in Cognition (ABC) will contain decision-making tasks with and without initial decisions. This will allow information search following an initial decision to be compared to information search without an initial decision. The proportion of confirming information selected in both types of decision-tasks will be compared for statistical significance to answer research question one.

The importance of subjective information assessment and selection are the current standards in measuring the presence of a confirmation bias during information acquisition and decision-making [4], [28], [5]. Associating additional behavior patterns with the standard information selection behavior allows for robust bias detection. Using models to mimic and predict the patterns of behavior associated with confirmation bias allows for a

quantitative understanding of what techniques and behaviors are associated with the presence of a confirmation bias. In light of these considerations, the alluring benefits of creating such models leads to the ensuing investigative question and hypothesis:

> **Research Question 2**: What are the information acquisition behavior patterns associated with a confirmation bias?
>
> *Hypothesis: Behavior patterns associated with a confirmation bias will be revealed by associating biased information selection* [10] *or information/question importance* [6] *with completion time, and information revisit.*

To assess research question two, the behavior patterns in information search time and information revisit will be associated with biased information search or information/question importance. If there are apparent trends between a specific behavior pattern and biased information search or information/question importance, statistical significance will be tested to determine if an information acquisition behavior can be associated with confirmation bias.

In determining the feasibility of bias detection from physiological signals, the performance of machine learning models for modeling such relationships is paramount. A model with high performance is desirable in predicting bias from physiological signals. This yields the following question and hypothesis:

> **Research Question 3:** Can a machine learning classification model using physiological signals estimate the presence of confirming and disconfirming information with performance greater than random chance?

Research Objective: Develop a machine learning model able to classify the presence of confirming information with equal-class-weighted classification accuracy greater than 50%.

To assess research question three, multiple machine learning models will be developed and tested for balanced accuracy greater than 50%. Although models with a balanced accuracy of 50% is not reliable enough for operational use, this elementary objective was used to determine if modeling the relationship between electroencephalography signals and biased behavior is possible.

To understand the underlying neurophysiological mechanisms associated with subjective information value and information selection, metrics will be mapped to neurophysiological measurements. These mappings will allow neurophysiological patterns associated with confirming and disconfirming information to be identified. The suitability of applying neurophysiological techniques in this domain are demonstrated by Minas et al.'s neurophysiological work which associated the activation of the brain's right frontal lobe with the presence of preference-supporting information. These results yield the following question and hypothesis:

Research Question 4: Are neurophysiological signals in the right frontal lobe associated with confirming and disconfirming information? Are neurophysiological signals in the right frontal lobe salient features in a machine learning information classification model?

Hypothesis: In contrast to disconfirming information, confirming information will provoke increased activity in the brain's right frontal lobe [10] *which will be significantly different in neurophysiological signals. The difference in activity will*

result in features associated with the brain's right frontal lobe being salient machine learning features.

To assess research question four, cross-participant time series signals of confirming and disconfirming information will be compared for statistical difference. In addition, salient features associated with the brain's right frontal lobe in the random forest models will be explored.

3.3 Experiment

The decision-making experiment will require the participant to complete a modified MITRE-developed Assessment of Biases in Cognition (ABC) [42] test while physiological measurements are collected. The ABC tests for behavior elicitation of numerous biases. This experiment uses a modified version of the ABC assessment which only contains behavior-elicitation tasks associated with confirmation bias. In addition, half the decision tasks included an initial decision while half did not (see Section 3.3.1.1 for justification details).

ABC elicits confirmation bias behavior in two paradigms: 1) information search decision making, and 2) evaluation/weighting of evidence/questions. The information search decision making paradigm is developed from research studies on confirmation bias in information search behavior [23], [43]. Tasks in this paradigm follow a three-step approach:

1) Participant makes and records an initial decision,

2) Participant selects additional information and,

3) Participant makes and records a final decision.

The additional information presented consists of confirming and disconfirming information relative to the participant's initial decision. Confirmation bias is quantified by the proportion of selected confirming information.

ABC utilizes a fictitious snack stand decision-making task to elicit biased behavior. In the snack stand task, participants must choose between opening a snack stand that sells either organic snacks or diet snacks, as illustrated in Figure 2. After assessing an initial decision, the participant is presented with 8 information headings; half support organic snacks and half support diet snacks as seen in Figure 3. The participant chooses information they would like to obtain more information on to make a final decision. Figure 4 illustrates a piece of information selected by the participant. After completing the information search, participants make a final decision, which concludes the task. ABC contains two additional fictitious "Stand" scenarios in which the decision is on opening two different types of bakery and two different types of exercise classes.

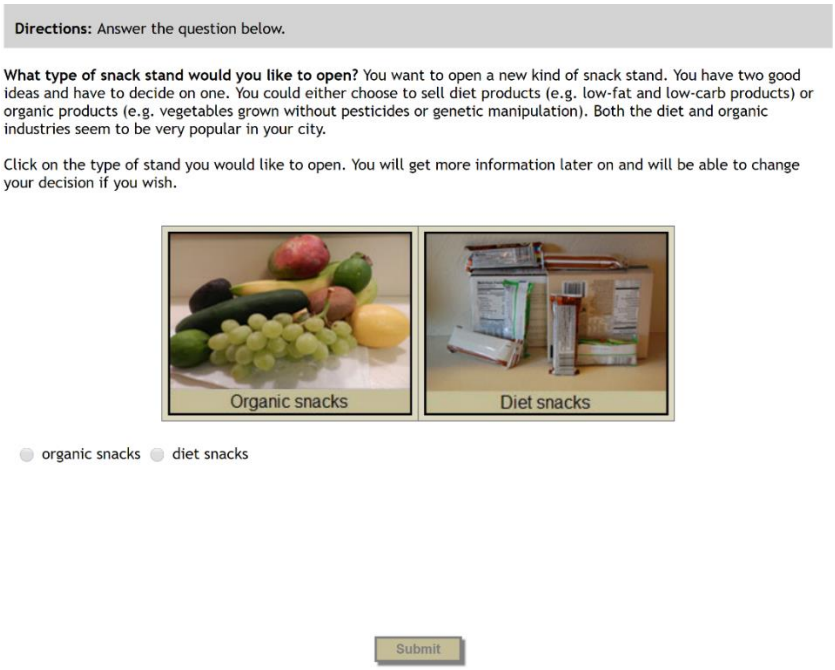

Figure 2: Snack Stand Decision Making Task

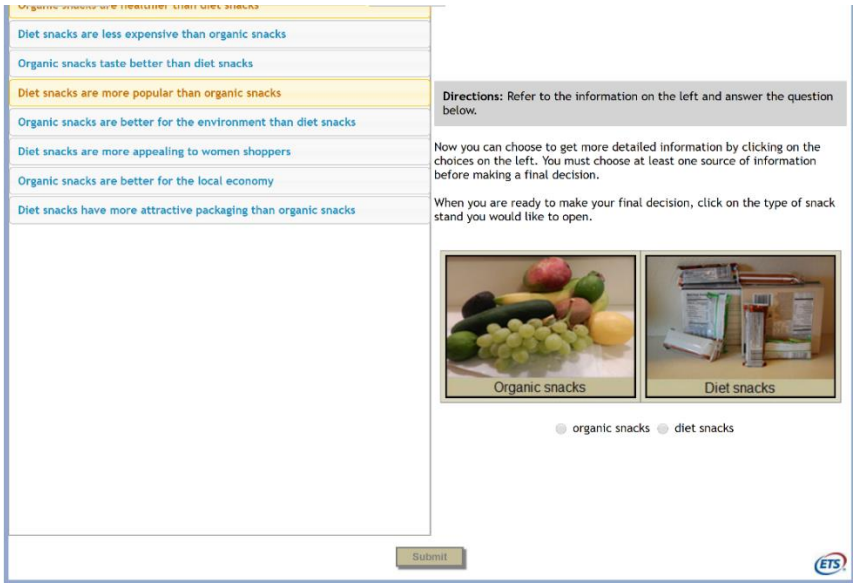

Figure 3: Snack Stand Information Search

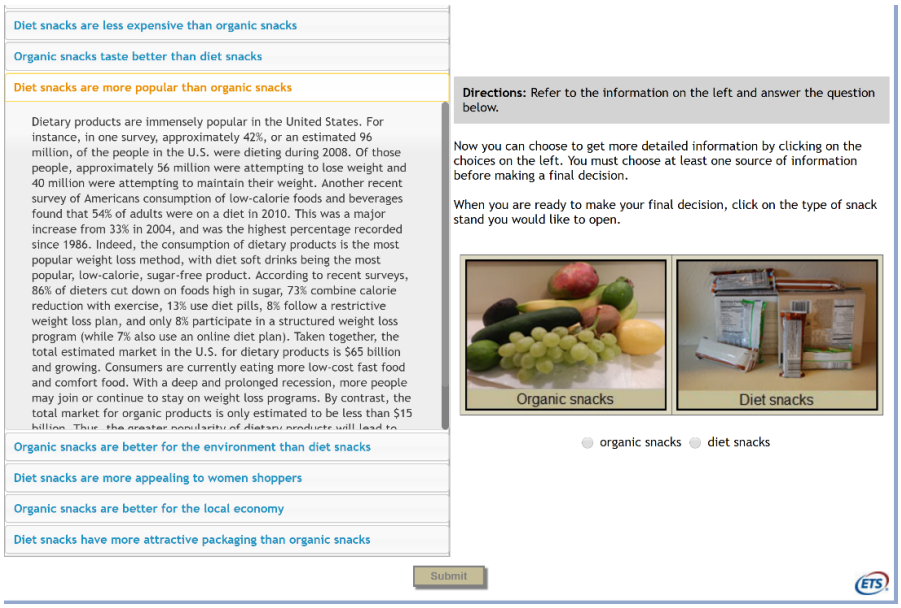

Figure 4: Snack Stand Selected Information

ABC also utilizes a second type of task in the information search decision-making paradigm known as a "Comparison" task. For this "Comparison" task, participants compare and choose between two products. In example, as seen in

Figure **5**, the participant must choose between two types of cars. After assessing an initial decision, the participant is presented with an unbalanced set of product comments. The comments are unbalanced by having more comments that support the participant's initial product decision. The participant must click on a comments header to display the comment. The comments position on each product is indicated by a green thumb up or red thumbs down (illustrated in Figure 6). To encourage effortless mental reactions, a fake monetary incentive and time limit is implemented [42]. Each comment the participant selects costs $1, which incentivizes only selecting comments deemed most important. These pressures are to increase the type of thinking that is conducive to biased behavior [42]. The level of confirmation bias in each task is quantified by the proportion

of selected confirming comments. ABC implements three additional product "Comparison" decision tasks of the same form which include choosing between types of cruises, music festivals, and gyms.

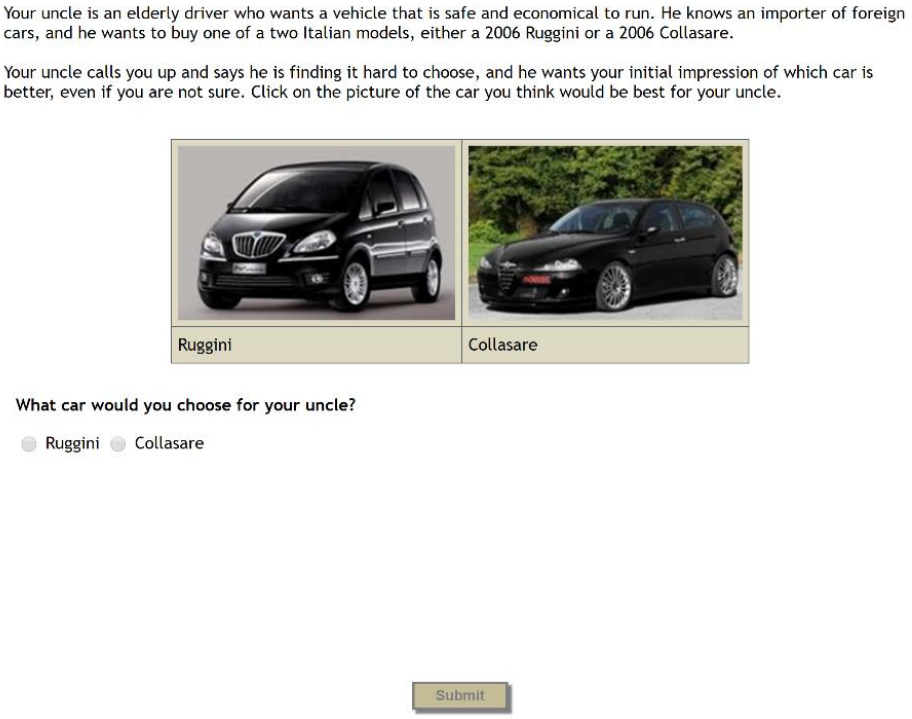

Figure 5: Car Comparison Decision Making Task

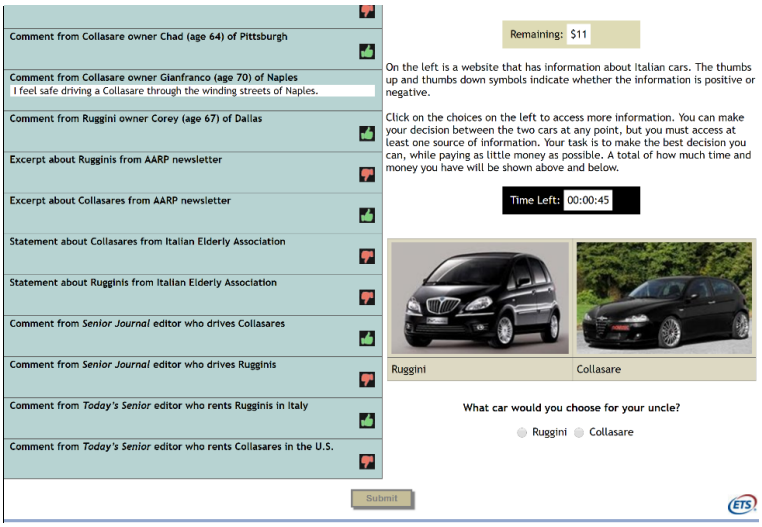

Figure 6: Car Comparison Information Search

The evaluation/weighting of evidence/questions paradigm is developed from research studies on confirmation bias in selecting questions or evidence to evaluate a specified hypothesis. Findings indicate people tend to select questions that would confirm the hypothesis when experiencing a confirmation bias [20]. Tasks in this paradigm present a fictitious scenario with a question and a hypothesis. The participant is presented with a list of evidence/questions to choose from; half of which confirm the hypothesis while the other half disconfirm. The participant chooses which evidence/questions are most important to answer the question presented in the scenario. A confirmation bias is quantified by selecting more confirming evidence/questions than disconfirming. ABC presents two scenarios in this paradigm. In one, the participant assumes the role of an intelligence analyst (Figure 7) and must select evidence that is most important to investigate to properly answer the question. The next scenario, the participant assumes the role of a human resources employee (Figure 8). The participant is tasked to select questions to assess a job applicant's specific attributes. Half of the questions confirm the attribute of concern, while the other half disconfirm the attribute. ABC contains three intelligence analyst tasks and four human resources tasks.

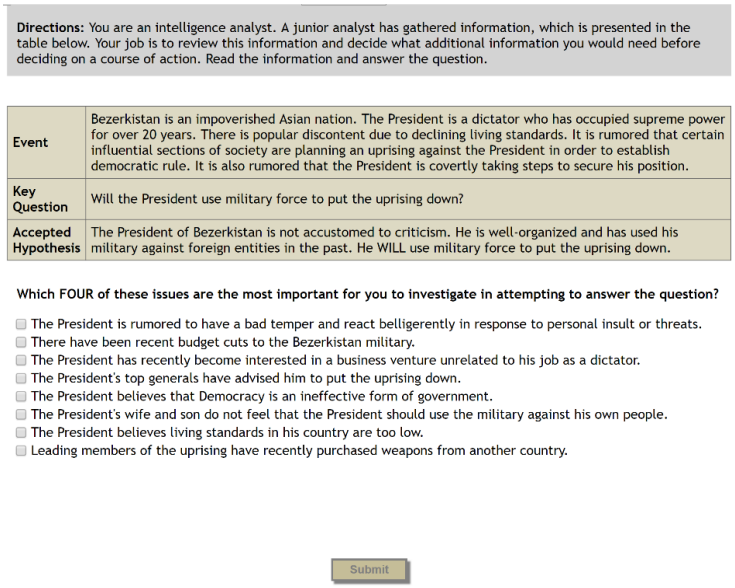

Figure 7: Intelligence Analyst Evaluation of Evidence Paradigm

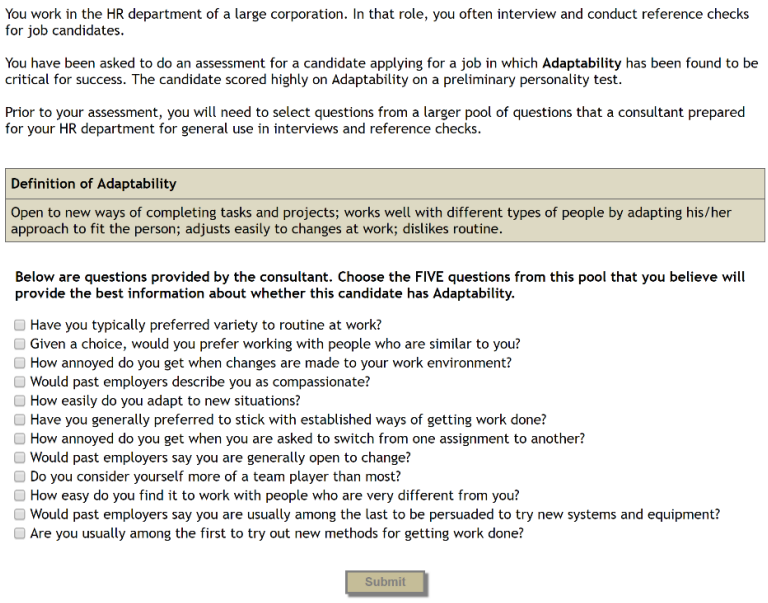

Figure 8: Human Resources Evaluation of Questions Paradigm

The complete outline of ABC tasks and their respective paradigm implemented in this experiment are in Table 1.

Table 1. ABC Tasks and Confirmation Bias Paradigm

Task	Paradigm*	Task Quantity	Task Versions
Stand	ISDM	3	Snack, Bakery, Exercise Class
Comparison	ISDM	4	Cruises, Music Festival, Working Out
Intelligence Analyst	EWEQ	3	Three events in intelligence analyst role
Human Resources	EWEQ	4	Four tasks in human resources role

*ISDM: Information Search Decision Making, EWEQ: Evaluation/Weighting of Evidence/Questions

3.3.1 Variables

3.3.1.1 Independent Variables

Both biased and unbiased decisions are desired for a robust model of the relationship between physiological data and confirmation bias. The variability of presence of a belief primer (initial decision/accepted hypothesis) in decision tasks will be used to create biased and unbiased decisions. In the Stand and Comparison tasks, approximately half of the tasks will have no initial decision. In the Intelligence analysis tasks, two of the three tasks will not have an accepted hypothesis.

Another independent variable implemented in the ABC test is the amount of confirming/disconfirming information present. In the product Comparison tasks, more confirming information is present than disconfirming information, while all other tasks have equal amounts of confirming and disconfirming information. The unbalanced information was shown to induce a biased behavior [42]. The independent variables with varying levels in the experiment are listed in Table 2.

Table 2: Independent Variable Summary

Independent Variable	Measurement precision	Settings	Predicted effects
Initial Decision	Present or Absent	[Present, Absent]	Absent = less biased information search
Accepted Hypothesis	Present or Absent	[Present, Absent]	Absent = less biased information value
Amount of confirming/disconfirming information	0-100% ratio of confirming/disconfirming	[50, 80]	Higher ratio of confirming information = more biased information search

3.3.1.2 Response Variables

Information selection and questions/evidence importance are response variables used to quantify the presence of confirmation bias. The association of the physiological response variables with the presence of confirmation bias is the objective of this experiment. The response variables in this study are categorized into the following three groups: 1) presence of a confirmation bias 2) behavioral patterns and 3) physiological signals

The presence of a bias can be indicated by the participant's behavior, as observed through selection of information or importance of evidence/questions. When a participant selects more confirming than disconfirming information during the information search phase of a decision task, the decision is labeled as biased. Similarly, when a participant values more confirming evidence or questions than disconfirming the decision task is biased.

Behavior patterns may also be different for participants experiencing a bias than those who are not. The response variables concerning behavioral patterns are: information search time, information revisits, information selection, and

evidence/question importance. Information search time is the time participants spend performing the information search portion of a task and may be indicative of a confirmation bias if only confirming information is sought. Information revisits is the cumulative number of times information is revisited during information selection of a task. This behavior could show uncertainty in the information selected and may be associated with confirmation bias. Information selection is quantified by the number of confirming and disconfirming pieces of information selected. Throughout this work, information selection is quantified by the proportion of selected confirming information to selected information. Lastly, evidence/question importance is the proportion of selected confirming evidence/question to selected evidence/questions. All of these behaviors may be different when the participant is experiencing a confirmation bias.

Physiological signals include EEG, ECG and EOG. EEG includes physiological measurements taken over the entire head. Correlation of subject information selection (confirming, disconfirming) with patterns in EEG measurements will be used to determine associated electrophysiological responses to confirming and disconfirming information. These correlations may be used to detect and quantify the presence of a bias. ECG includes physiological measurements across the chest and capture heart rate. EOG includes eye movement and blinks which can be used to measure attention as well as for removing eye artifacts from EEG. All collected response variable for this experiment are outlined in Table 3.

Table 3. Response Variables

Response variable	Normal Operating Level and Range	Measurement Precision	Relationship to objective
Information Selected (categorical)	["confirming", "disconfirming]	1 piece of information	Degree of confirmation bias
Question/Evidence Importance (categorical)	["confirming", "disconfirming"]	1 question /piece of evidence	Degree of confirmation bias
Initial Decision (categorical)	[Choice 1, Choice 2]	Subjective	Establish belief
Final Decision (categorical)	[Choice 1, Choice 2]	Subjective	Decision Task
EEG (numerical)	0-131 Hz at 500 samples/sec	0.7 μV RMS from 1-50 Hz	Delta: < 6 Hz Theta: 7-11 Hz Alpha: 12-15 Hz Beta: 16-22 Hz Gamma: 22-30 Hz
ECG (numerical)	60-100 beats/min at rest	Beats/min	Stress/workload
EOG (numerical)	Mean = 17 blinks/minute	0.7 μV RMS from 1-50 Hz	Movement, visual attention
Information Search Time (numerical)	0-500 seconds	1 second	Behavior
Mouse Clicks (numerical)	1-20 clicks	1 click	Behavior
Information Fixation Time (numerical)	0-500 seconds	1 second	Behavior

3.3.1.3 Constant Factors

Within the ABC test, constant factors include the structure of the test and the information presented for tasks. Each participant will take the same ABC test. The test is anticipated to last no more than 60 minutes for completion. This short test period is expected to mitigate the effects of fatigue on the participants decision-making behaviors. In addition, each task will have confirming and disconfirming information present. This

allows a confirmation bias to be quantified. A summary of the factors that will remain constant are outlined in Table 4.

Table 4. Constant Factors Summary

Factor	Desired experimental level	How controlled?	Anticipated effects?
Task Count	14 Decision Tasks	Standard ABC test format	Mitigate fatigue/balanced distribution of biased decisions
Information Type	Presence of both confirming/disconfirming information	Standard ABC test format	Despite the presence of both types of information, a bias will occur

3.3.1.4 Nuisance (Confounding) Factors

In human-subject experiments, many uncontrolled factors can influence results. Human-subjects are naturally complex and highly variable. Table 5 outlines the expected nuisance factors and the strategy of mitigation.

Table 5. Nuisance Factor Summary

Nuisance Factor	Strategy	Anticipated Effects
Learning Effect: Test progression may result in decreasing information search time, which would make information search time a poor behavioral measure.	Instruction are provided with each decision task and the tasks do not require outside knowledge to complete.	A minor decrease in information search time may be exhibited by participants due to task familiarity, but because of the simple nature of the decision tasks the time will be negligible.
Misinterpretation of instructions: Confusion on tasks could lead to undesired brain activity and false measure of confirmation bias.	Instruct participant to read instructions clearly prior to beginning each task.	Some participants may not follow instructions, but a majority of participants will read and follow instructions.
Unbalanced distribution of biased and unbiased decisions will make associating physiological signals with bias difficult.	Test is composed of tasks with primers and without primers.	Tasks with primers will cause a unbalanced information search while task without primers will cause a balanced information search.
Participants external belief reflected in tasks: If participant does not believe either of the available decisions, all information would be disconfirming and result in erroneous results.	Prime participants belief by making initial decision or presenting accepted hypothesis	Some participants will likely have external beliefs reflected in the decision process, but most participants have a decision option which reflects their belief.

3.3.1.5 Known/Suspected Interactions

During the decision-making process, people tend to make estimates based on initial values. In this experiment, this estimate is the initial decision made prior to information search. Anchoring bias occurs when people anchor on these initial estimates and fail to properly adjust their decision [25]. Given participants may anchor on their initial decision, there will likely be anchoring bias in addition to confirmation bias present in the tested decision-making tasks.

3.3.1.6 Test Matrix

tasks.

Table 6 outlines the test matrix for the experiment used for all participants. Each version of the tasks (Stand, Comparison, Evaluate Evidence, and Evaluate Questions) is distributed throughout the test to maintain participants' focus. In addition, the presence of a primer is distributed across the tasks. Lastly, it is important to note that while there are only nine tasks listed, the intelligence analyst (Intel) and human resources (HR) tasks have three and four subtasks respectively. Including the subtasks, the ABC test consists of 14 tasks.

Table 6: Test Matrix

Task	Form (version)	Primer	Degree of Confirmation Bias
1	Stand (Snack)	Present	High
2	Comparison (Car)	Absent	Low
3	Evaluate Evidence (Intel-Analyst)	Present (1/3 subtasks)	High/Low
4	Stand (Bakery)	Present	High
5	Comparison (Music)	Absent	Low
6	Evaluate Question (HR-Dept)	Present	High
7	Comparison (Working out)	Present	High
8	Stand (Exercise)	Absent	Low
9	Comparison (Cruises)	Present	High

3.3.2 Participants

Fifteen United States military and government civilian personnel participated in the study. All participants were male with ages between 21 and 51 with a mean age of 29.4, median of 28 and standard deviation of 7.28. All participants had a Bachelor's Degree or higher and used computers daily in their job. With the exception of one outlier, all participants obtained "fair" or better sleep quality of 5 to 9 hours. The one outlier

obtained 0 to 4 hours of "very poor" quality sleep prior to participation in the study. Inclusion criteria included the ability to operate a computer mouse, be at least 18 years of age, and be a US citizen. Exclusion criteria included visual impairments preventing viewing a computer screen, the inability to operate a computer mouse, and specific motor perceptual, or cognitive conditions that preclude operating a computer. Prior to beginning the experiment, participant consent was obtained. Due to placement location of the ECG sensors, additional participant consent was collected and the participants were allowed to self-apply the sensors if they chose. Participants did not receive compensation for their participation in the study.

3.3.3 Materials

The ABC Test software used in this experiment is a modified version of the MITRE-developed ABC Test Delivery Platform [42]. The original software was modified to allow collected physiological data to be time stamp marked according to specific events during test administration. These event makers in the physiological data allow for proper post processing to correlate test events with physiological signals. The modifications otherwise have no impact on the test interface or implementation which allows this experiment to benefit from the extensively tested ABC test interface [42]. Aside from the two computers required for test administration and physiological data collection, other necessary equipment includes the three physiological sensors, which are covered in further detail in the ensuing sections.

3.3.3.1 ABC Test

The ABC Test is administered through a web browser. The test delivery platform requires PostgreSQL 9.3 database and Apache Tomcat 8 server to be installed. Specific software configuration details can be found in the ABC User Manual [44]. The database contains the test content and stores all collected data when the test is administered. The ABC test software Web Application Resource (WAR) file resides on the Tomcat server and is accessible through a web browser by navigating to "localhost:8080/stiEditor". Figure 9 shows the notional software configuration on test administration computer. Although remote test administration of the ABC test is possible, the collection of physiological data requires the test to be administered on the computer that the server and database resides. For this experiment, the ABC software was modified to connect to the Cognionics trigger hardware through USB on the computer the software resides. Further details on events marked in the physiological data through the trigger hardware is provided in Section 3.3.3.1.2.

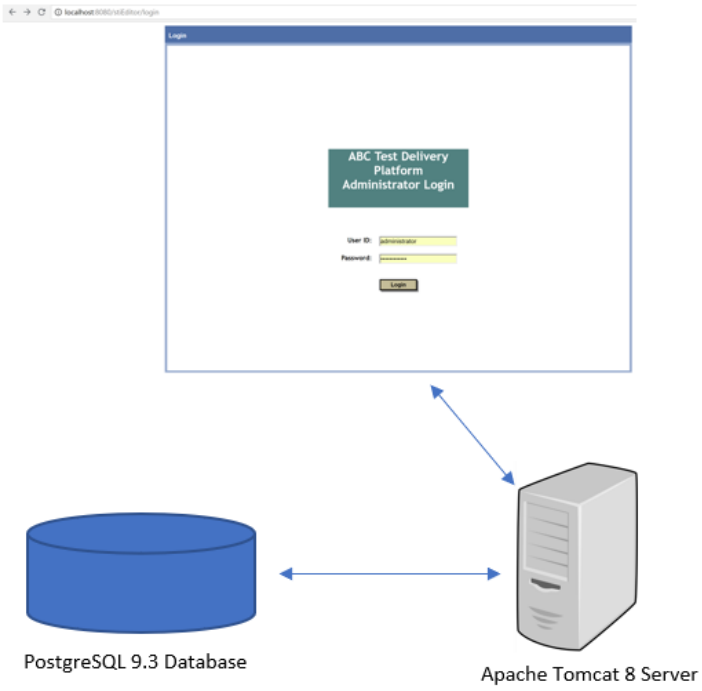

Figure 9: Notional Software Configuration of Test Administration Computer

3.3.3.1.1 Interface

After the test administrator entered the participant ID, the participant began the test be clicking the "Begin" button. Each page in the test has instructions for the respective task. The "Submit" button is greyed out and disabled until the task is completed. All answers in the test are submitted with mouse clicks through radio buttons or check boxes. As the participant completes the test, their progress is indicated in the top left corner of the interface. Their progress is indicated by the current test number out of the total tests i.e. Test 1 of 9 as shown in Figure 10. The "Test" progress is synonymous with this works "Tasks" in the Table 6. Additionally, the Test Time Elapsed for the entire ABC test is displayed in the top right corner of the interface.

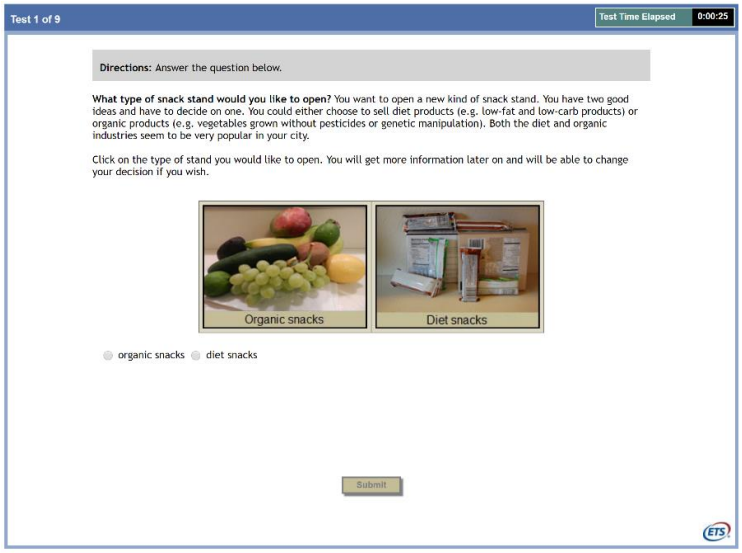

Figure 10: ABC Test Interface

3.3.3.1.2 Timing Database and Triggers

The collection of physiological signals while the ABC test is administered requires a two-computer set-up. The ABC test is administered on the assessment computer while the physiological measurements are collected and logged on the physiological collection computer as illustrated in Figure 11. The assessment computer is connected to trigger hardware through USB, which communicates wirelessly with the data acquisition unit (DAQ). The triggers sent from the Assessment computer mark the EEG data to associate events with EEG signals in post-processing. Events in the ABC test that send a marker through the trigger include test begin, page load, mouse click, submit answer and test end. The DAQ is physically connected to the EEG cap on the participant and wirelessly sends triggers from the Assessment computer and EEG signals from the EEG cap to the physiological collection computer. The physiological collection

Collection computer also receives other collected physiological signals like ECG and EOG through a physically connected Auxiliary Module.

Figure 11: Two Computer Physiological Measurement Collection Set-up

3.3.3.2 Physiological Recording Devices.

The Cognionics Mobile-72 system was used for all physiological data collection. The Mobile-72 system collects 64 EEG voltage channels as well as 8 auxiliary channels for ECG and EOG. The Cognionics EEG cap shown in Figure 12 is fitted onto the participant to collect brain activity through EEG. The 64 EEG electrode locations are based on the International EEG 10-20 electrode placement. For proper EEG data collection, the electrodes on the cap must have high conductance with the participant's head. To increase conductance, gel is applied to the electrodes until the impedances are below 100K ohms.

Figure 12: Cognionics EEG Cap

The vertical and horizontal EOG electrode locations are shown in
Figure 13 and the ECG electrode locations are shown in Figure 14. The experiment administrator attaches the EOG electrodes. For privacy reasons, the participant is instructed on ECG electrode placement and attaches electrodes in a private room if desired.

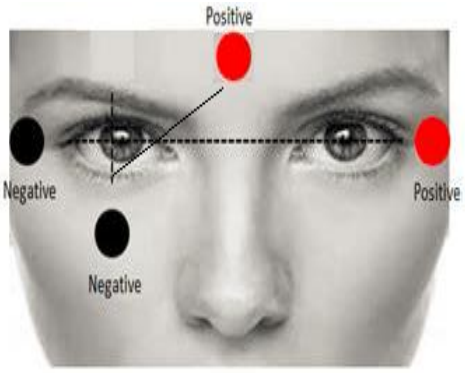

Figure 13: Vertical and Horizontal Electrooculography Node Locations

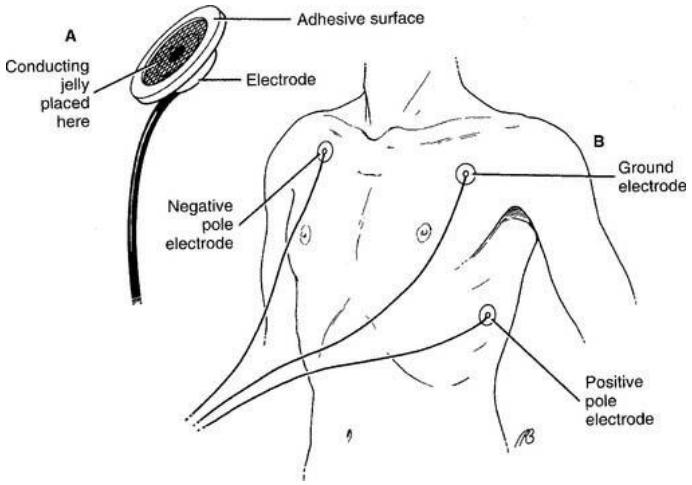

Figure 14: 3-Lead Electrocardiograph (ECG) Node Locations

3.3.4 Procedures

The pre-experiment procedure includes conducting the informed consent briefing/completing the informed consent document (ICD), assigning a participant number, measuring the participants head size for proper EEG cap sizing and scheduling the experiment session, as shown in Figure 15. The experiment session was no more than two hours in length to prevent participant fatigue. The period includes behavioral and physiological signal measurements of participants while they complete the assessment. Pre- and post-testing questionnaires were given to each participant on the experiment day.

Figure 15: Pre-experiment Procedure

Figure 16: Testing Session Sequence

Each participant individually completed the testing day activity sequence (Figure 16). Participants were scheduled at different times, in 2-hour blocks to allow for adequate preparation, task completion, and cleanup. Upon completion of the experiment, participants were asked to not discuss the nature of the task to prevent bias induction in participants who had not completed the experiment.

3.3.5 Data Collection

ABC response data collection is completed using a PostgreSQL database. The ABC interface exports each participant's test data from the database into a .csv file. Recorded data includes the following events with time stamps: test start, page load, answer and mouse clicks as illustrated in Figure 17.

Test ID	Participant ID	Timestamp	Event Type	Object ID	Item	Item Name	Page	Question	Question Type	Response
60	0000	0	TEST_START		0		0			
60	0000	0	PAGE_LOAD		0		0			
60	0000	134779	PAGE_LOAD		1		1			
60	0000	136924	PAGE_LOAD		2		1			
60	0000	152403	MOUSE_CLICK	1	2		2	1	SingleChoiceOption	1
60	0000	153619	ANSWER	98857	2	Snack-Stand-BE6	2	1	singleChoice	1
60	0000	153736	PAGE_LOAD		2		2			
60	0000	185163	MOUSE_CLICK	1	2		3	1	SingleChoiceOption	1
60	0000	186098	ANSWER	98859	2	Snack-Stand-BE6	3	1	singleChoice	1
60	0000	186184	PAGE_LOAD		2		3			
60	0000	194419	MOUSE_CLICK	1	2		4	1	InfoItem	1
60	0000	206731	MOUSE_CLICK	5	2		4	1	InfoItem	5
60	0000	220307	MOUSE_CLICK	3	2		4	1	InfoItem	3
60	0000	223122	MOUSE_CLICK	4	2		4	1	InfoItem	4
60	0000	243426	MOUSE_CLICK	1	2		4	2	SingleChoiceOption	1
60	0000	244779	ANSWER	8771;8770;	2	Snack-Stand-BE6	4	1	infoItemAccordion	5;4;3;1
60	0000	244779	ANSWER	98861	2	Snack-Stand-BE6	4	2	singleChoice	1
60	0000	244886	PAGE_LOAD		2		4			

Figure 17: Example Raw Data Output from ABC Test

Each event includes attributes which indicate the current task, question, and item for the respective event. All electrophysiological signals are collected by Cognionics Data Acquisition Software and saved in the Biosemi (.BDF) file format. Unique triggers mark the electrophysiological time-series data at the time of important stimulus onset or participant actions. In Figure 18, the markers indicate a single choice (5376), answer submission (4352) and page load (512).

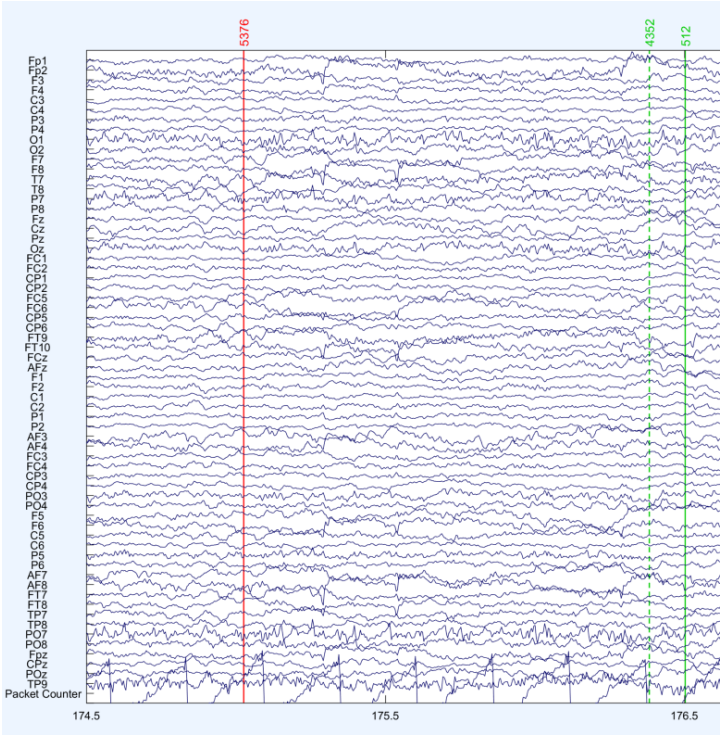

Figure 18: Example EEG Data Plot with Trigger Markers

3.3.6 Analysis Strategy

Data collected from the ABC test will be analyzed using statistical packages in Python. Machine learning analysis of collected physiological data is covered in the machine learning pipeline section of this chapter (Section 3.4). The main objectives for analysis of the ABC test data are to 1) quantify degree of confirmation bias for each task, 2) analyze behavior patterns associated with confirmation bias, and 3) assess if physiological signal activity in the brain's right frontal lobe is associated with confirming information.

The degree of confirmation bias in each task is quantified by calculating the proportion of selected confirming to disconfirming information/questions. A proportion greater than 0.5 indicates confirmation bias while a proportion less than 0.5 indicates

unbiased tasks. A proportion closer to one indicates a higher degree of confirmation bias. To analyze the effect of an initial decision on subsequent information search, a Wilcoxon Signed-Rank test will be used. This will test for a difference in the proportion of confirming information in tasks with an initial decision compared to tasks without an initial decision. The Wilcoxon Signed-Ranks test will be utilized because each participant will complete tasks with and without an initial decision and the data is not necessarily normally distributed. Each decision-task with an initial decision will be compared to decision tasks without an initial decision. The paired input will be a participant's proportion of confirming information in each task. The statistical significance for these tests will be critical to answering research question 1 (Section 3.2).

As a data exploration step, the behavioral data will be plotted using histograms to determine if there are any trends associated with bias. If there are any apparent trends in the behavioral data, regression models will be created to model the relationship between information revisit, information search time, and degree of confirmation bias. A notional graph illustrating a possible linear relationship between task completion time and degree of confirmation bias is shown in Figure 19. If a similar relationship between response time or information revisit and degree of confirmation bias appears in the data, these behaviors will be implemented in the machine learning models for predicting bias.

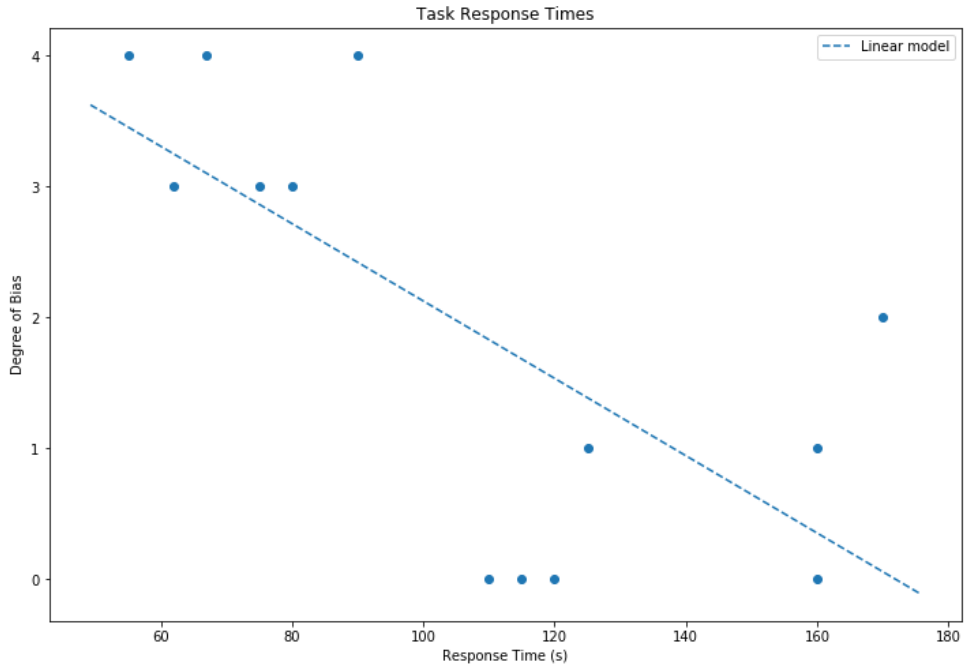

Figure 19: Notional Linear Model of Response Time and Confirmation Bias

Lastly, time series signals across participants will be compared to assess if increased activity in the brain's right frontal lobe is associated with confirming information. To analyze the difference between the two information conditions, a nonparametric statistical test using Monte Carlo permutations with cluster corrections in EEGLab will be performed on the time series signals of confirming information and disconfirming information across all participants [45]. EEG channels associated with the brain's right frontal lobe, specifically F2, F4, F6 and F8, at the time interval of 100ms to 400ms following stimulus onset will be explored [10].

3.4 Machine Learning Pipeline

This section outlines the end-to-end supervised machine learning (ML) pipeline employed to analyze and create models from collected physiological data. Unless

otherwise specified, all the steps in the machine learning pipeline are within-participant. For this work, within-participant means a model is trained on data from a single participant and the performance metrics are measured using data from that same participant. The diagram in Figure 20 is a modified version of Raschka's pipeline [46] and provides a visual representation of the pipeline used in this methodology. The pipeline begins with the data preprocessing steps necessary to transform the collected raw EEG data into an ML-ready format. Next, cross-validation is implemented as a sampling technique. After data preparation, models are trained on each cross-validation fold and performance metrics are analyzed. One important aspect of this applied pipeline is all performance metrics are cross-validation metrics because of the dataset sizes. For this reason, there was no hyper-parameter optimization or test dataset. The ensuing sections cover the technical details of each step in the pipeline in further detail.

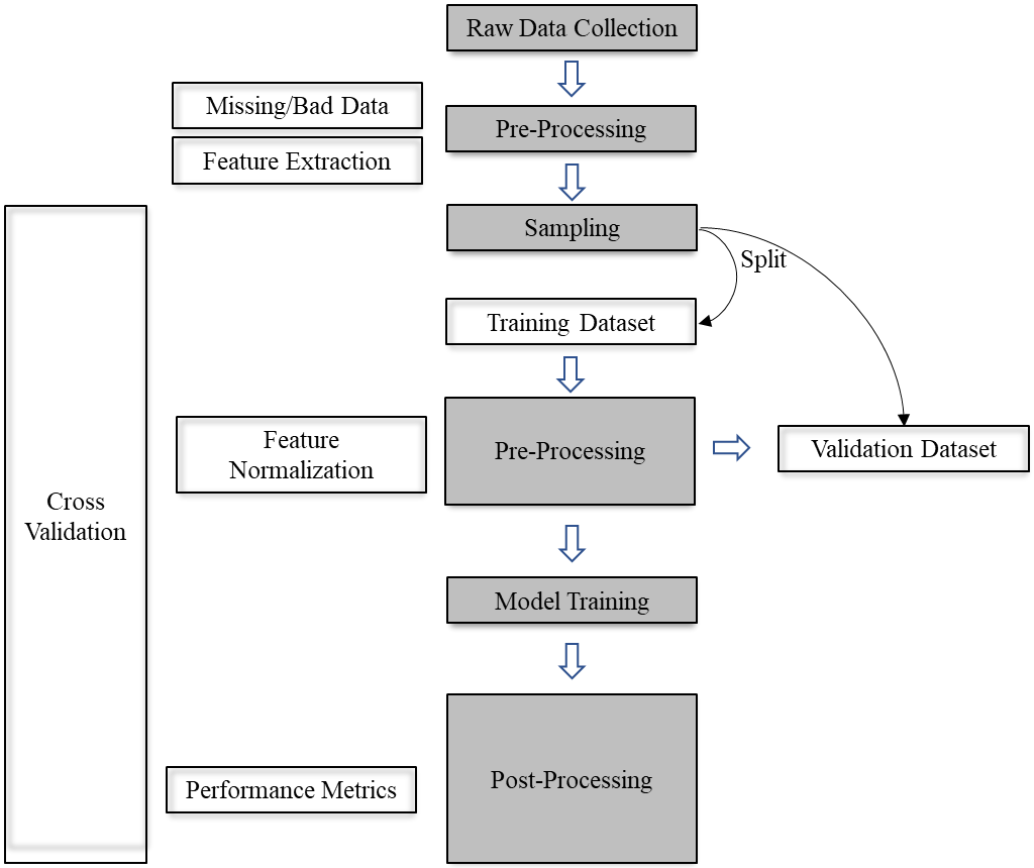

Figure 20: Machine Learning Pipeline

3.4.1 Data Preprocessing

Upon completion of the experiment in Section 3.3, the collected raw physiological data was in a BDF file format. During EEG collection, the data acquisition device was observed to sporadically malfunction. Given this was a pilot study, the data was still utilized. The noise due to the malfunctioning device was removed to the furthest extent possible, but may still result in poor EEG data. The raw EEG data of each participant was cleaned using EEGLab version 14.1.1 [47] according to Makoto's preprocessing pipeline, [48]. The preprocessing pipeline performed on the raw EEG data of each participant included the following steps:

- Down sample data from original 500 Hz sampling frequency to 250 Hz.
- High-pass filter at 1 Hz to remove baseline drift with a basic finite impulse response (FIR) filter.
- Import International 10-20 system channel information.
- Remove line noise with CleanLine EEGLab plugin.
- Reject bad channels using Automatic channel rejection using kurtosis with a Z-score threshold max of 5.
- Interpolate removed channels to prevent bias during referencing to average.
- Re-reference data to the average.
- Perform Independent Component Analysis (ICA) to determine eye-blink components. Adjust rank to account for removed channels.
- Remove components associated with eye-blink artifacts using icablinkmetrics plugin with the vertical EOG channel for eye-blink artifact reference [49].

After preprocessing, the data was segmented in EEGLab using two different approaches: 1) segmentation by task, 2) segmentation by information selection. The first method of data segmentation was to segment the data by task. In this approach, the EEG data corresponding to the information search portion of each of the 14 tasks was separated into a new file. Segmenting by task yielded one file for each individual task producing 14 separate files per participant.

The second method of data segmentation was to segment the data by information into epochs. When segmenting the data by information selection, each epoch was a two-

second window of data centered on when the participant selected information throughout the entire assessment. Any overlapping epochs were discarded from the dataset. For this method of segmentation, information selection across all tasks were epoched. Segmenting by information selection produced one file per participant.

Following segmentation, noisy data was rejected by visual inspection in EEGLab. Epochs with a large amount of noise relative to the entire dataset were rejected. An example of rejected noisy data is shown in Figure 21.

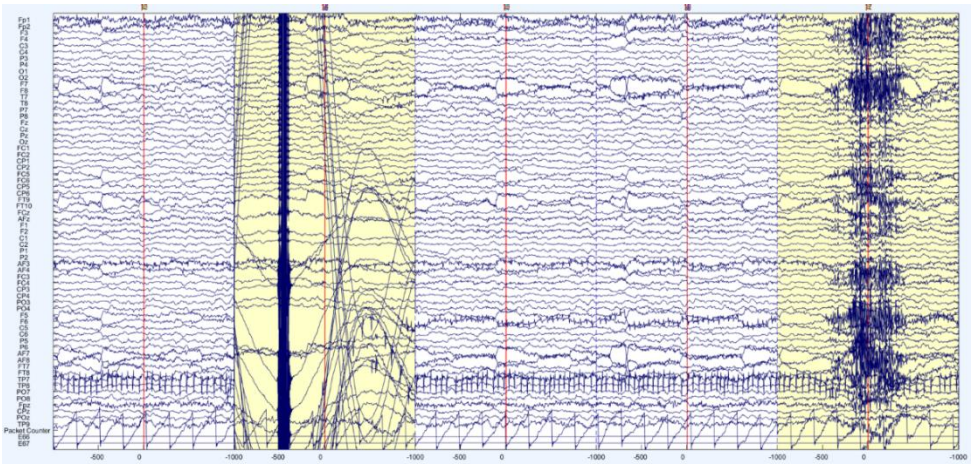

Figure 21: Two Visually Rejected Epochs in a Segment of EEG Data.

3.4.1.1 Time Series Feature Extraction

When segmenting the data by task, one file contains the information search portion of a single task. For example, if a participant spent 120 seconds on the information search of the first task, the respective file contains 120 seconds of EEG data. To prepare each segmented task for machine learning, a sliding window extracted smaller sequences from the larger segment. The implemented sliding window has two adjustable parameters: window size and step. The window size is the size of sequence extracted

from the larger segment. The step is the time points the window moves to extract the next sequence. For example, a task of 120 seconds at a sample rate of 250 Hz contains 30,000 frames. A sliding window with a window size of two seconds or 500 frames, and a step size of two seconds creates 60 two-second non-overlapping epochs from the 120 second task. This process created multiple epochs from one task. When segmenting the data by information selection, the segmentation process created two-second epochs, so a sliding window was not necessary. For time series classification, each epoch is an observation and each observation has 64 features which correspond to the EEG electrodes.

3.4.1.2 Frequency Feature Extraction

Epochs were transformed to the time-frequency domain using MATLAB. A complex Morlet wavelet was implemented for the time-frequency transformation due to the transient nature of EEG time-series data. This frequency transformation resulted in each epoch having the power spectral density of the standard clinical frequency bands: delta (1-6 Hz), theta (7-11 Hz), alpha (12-15 Hz), beta (16-22 Hz), and gamma (22-30 Hz) [29]. The mean of the power spectral density for each observation was then calculated. With mean power of the five frequency bands for a 64 electrode EEG cap this produced 320 features per subject. These features were inputs to the ML model as a single observation.

3.4.2 Datasets

Although there is originally one EEG recording per subject after collection, through the multiple data segmentation and feature extraction techniques there were four distinct datasets for ML:

1) Time Series Signal per Task

2) Frequency Features per Task

3) Time Series Signal per Information Selection

4) Frequency Features per Information Selection

The Time Series Signal per Task dataset consists of epochs from each task. Any individual epoch only contains data from one task. The label for each individual epoch is "biased" or "unbiased". If the epoch is from a decision-task in which the participant was biased, the label is 1, otherwise the label is 0 which is indicative of unbiased task; see Section 3.4.2.1 for details on how tasks are quantified as biased or unbiased. Because a participant was either biased or unbiased for an entire task, all epochs from the same task for any participant have the same label. For example, all epochs from a biased decision-task have the label 1, while all epochs from an unbiased decision-task have the label 0. The ML problem for this dataset is a many-to-one binary classification in which the goal is to classify a time series sequence of EEG data as biased or unbiased. The input has the shape (batch size, time steps, features). Batch size is variable and is the number of observations. Time steps is the number of frames in the epoch and is equal to the window size of the sliding window. Features is equal 64 and corresponds the number of EEG channels since the epoch is of a time series sequence of EEG data.

The Frequency Features per Task dataset is labeled in the same manner as the time series signal per task. The main difference between these two datasets is that instead of the dataset being time series sequences in each epoch, the dataset is the mean power of the clinical frequency bands of the time series sequence. The time-frequency transformation converts the time series sequence into a single mean value. The ML

problem for this dataset is one-to-one binary classification to classify an observation as biased or unbiased. The input shape into a ML model is (batch size, features). Batch size is variable and is the number of observations in the input. Features is the mean power of each of the five clinical frequency bands at each EEG channel. A single observation contains 320 features due to every channel in 64-channel EEG data having the mean power of each of the five clinical frequency bands.

The Time Series Signal per Information Selection dataset consists of time series data when a participant selected a piece of information during a decision-making task. Each epoch is labeled as “confirm” or “disconfirm”. A label of 1 indicates the information is confirming, while a 0 indicates the information was disconfirming, see Section 3.4.2.1 for further details on labeling information as confirming or disconfirming. The dataset consists of information selection across all decision-tasks. The ML problem for this dataset is a many-to-one binary classification in which the classification goal is to classify information as confirming or disconfirming. The input for this dataset has the shape (batch size, time steps, features) which is identical to the time series signal per task dataset.

The Frequency Features per Information Selection dataset is the frequency transformation of the time series EEG information selection dataset. This dataset is labeled in the same manner using “confirm” or “disconfirm” as labels. The ML problem for this dataset is a one-to-one binary classification of classifying a selected piece of information as confirm or disconfirm. The time-frequency transformation of the time series sequence converts the sequence to a single mean quantity. The input shape for this dataset is (batch size, features). Features is the mean power of each of the five frequency

bands across the epoch at each of the 64 EEG channels resulting in 320 features per observation.

3.4.2.1 Dataset Labels

There are two different "truth" labels for the datasets: bias and confirm. The bias label is a binary classification label for a two-class problem in which a one represents a biased task and a zero represents an unbiased task. The bias label was employed for datasets in which the data was segmented by task. Each individual task is labeled based on the proportion of confirming information the participant selected in the task. If the proportion of confirming information is greater than the proportion of disconfirming information, the entire task is labeled as biased or assigned a one, otherwise the task is labeled as unbiased or zero.

The confirm label is a binary label for a two-class problem in which a one represents the information confirms the participant's belief and a zero indicates the information disconfirms the participant's belief. The confirm label was employed for the datasets in which the data was segmented by information. When categorizing information in a task as confirming or disconfirming with respect to the participant's belief, the participant's belief must be known. Usually, the participant's initial decision is used as a way to tease out or establish the participant's belief [5], [6], [10]. Given one independent variable in the experiment design was to create a version of each task form without an initial decision to create a more balanced information search, the participant's final decision was used as the participant's belief for these tasks. Table 7 outlines the participant's decision used to establish the participant's belief in each task. For the

"Evaluate Evidence" and "Evaluate Questions" form of task, there was no initial or final decision, but rather the participant selected evidence or questions that were by nature confirming or disconfirming relative to the question/evidence in the task.

Table 7. Decision used to establish belief

Task	Form (version)	Decision used to establish belief
1	Stand (Snack)	Initial
2	Comparison (Car)	Final
3	Evaluate Evidence (Intel-Analyst)	Evidence Selected
4	Stand (Bakery)	Initial
5	Comparison (Music)	Final
6	Evaluate Questions (HR-Dept)	Questions Selected
7	Comparison (Working out)	Initial
8	Stand (Exercise)	Final
9	Comparison (Cruises)	Initial

In the "Stand" form of tasks, only confirming or disconfirming information was present. The two decision options were mutually exclusive such that confirming information for choice one was disconfirming for choice two, and vice versa. Only selected information was included when determining the proportion of confirming information.

For the "Comparison" form of tasks, there were two mutually exclusive choices for a decision and four types of information present: 1) supportive of participant's choice; 2) unsupportive of other choice; 3) supportive other choice; 4) unsupportive of participant's choice. The first two types of information were categorized as confirming since these types of information supported the participant's choice either directly or indirectly. The last two types of information were categorized as disconfirming. Only

selected information was included when determining the proportion of confirming information.

The “Evaluate Evidence” tasks present a key question and the participant selects four of the most important pieces of evidence in answering the question. Four pieces of evidence confirm the key question, while four disconfirm the key question. The proportion of confirming evidence was determined by the quantity of selected confirming evidence.

The “Evaluate Questions” tasks followed a similar approach except participants were presented with a definition of a personality trait, which was critical for success in a simulated job. The participant assumed the role of a human resources employee and selected five questions they believed would provide the best information about whether or not a candidate would have the desirable personality trait. The proportion of confirming questions was determined based on the quantity of confirming questions the participant selected out of the five total questions chosen by the participant.

3.4.2.2 Challenges of EEG

When applying ML to EEG data there are few important factors to consider. First, EEG signals have a poor signal-to-noise ratio: the data has significant amounts of noise and outliers [50]. Although the employed pre-processing steps aim to reduce such signal noise, each individual epoch still contains significant amounts of noise. A large number of observations is necessary to obtain high performance in ML due to this significant noise. Secondly, EEG signals are of high dimensionality because features can be extracted from multiple channels over different time steps [50]. The 320 features for the

datasets that use the five frequency bands demonstrate the high dimensionality. The frequency features per information selection dataset only contains 40-90 observations per participant, while the frequency feature per task dataset only contains 200-800 total observations per participant. For these datasets the curse of dimensionality will become a factor in training models. Necessary steps to prevent overtraining must be considered in the ML approach to the outlined EEG classification problems.

3.4.3 Approach

With the datasets in mind, there are two general ML problems on hand: many-to-one binary classification and one-to-one binary classification. The goal of the many-to-one binary classification problem is classifying a time series of EEG data. Depending on the dataset, the target variable for classifying the time series is either confirm or bias. Due to deep learning model's high performance in sequence classification, deep learning methods were used for many-to-one binary classification. The deep learning models which were evaluated for this task are long short-term memory (LSTM) and temporal convolutional networks (TCN) because of their near state-of-the-art performance in sequence classification [39]. For the one-to-one binary classification random forest (RF), linear discriminate analysis (LDA) and fully connected artificial neural networks (ANN) were explored. Multiple ML models were implemented for each dataset as an exploratory approach for the EEG analysis in this work.

3.4.3.1 Machine Learning Models

In light of random forest's ability to select the best features, all 320 features were used. Rather than tuning max features and max depth, set values were used to prevent

overfitting. Max depth was set to a value of 10 which was based on the training accuracy. Max features were set to be 10% of the total number of observations. The features were limited to reduce dimensionality because of the small dataset size. Max features for each participant varied and ranged from 4 to 9. Finally, each participant's model was trained on each cross-validation fold using these parameters with 500 trees.

LDA was utilized because of its stability with a small number of observations [33]. Generally, feature selection is implemented through an iterative approach to reduce the dimensionality for LDA but unfortunately this was not possible when reporting cross-validation metrics. To help account for the high dimensionality without selecting features, the LDA shrinkage parameter was set to 'auto' and the solver was set to 'lsqr'. These settings improve the LDA estimation of covariance matrices for datasets with high dimensionality [51].

A simple artificial neural network (ANN) with three fully connected layers was implemented. An ANN was selected because of its relative simplicity to other deep learning models. The ANN model design choices were selected through exploration by observing training accuracy with the goal of ensuring the models could adequately learn the training data. The model with the simplest architecture and lowest training time meeting the aforementioned goal was implemented for the final model. Optimizers which were explored included Stochastic Gradient Descent, RMSProp and Adam. Learning rates of 0.1, 0.01 and 0.001 were also explored. Layer units of 32, 64, 128, 256 and 512 with a depth of 1 to 5 were explored.

As outlined in Figure 22, the layers of the implemented ANN include a fully connected Dense layer with 512 units and a rectified linear unit (ReLu) activation

function followed by a dropout layer of 0.2. This sequence repeats two times. The input and output of each layer has the shape (None, 512). The final layer is a fully connected Dense layer with 1 unit and a sigmoid activation function which serves as the output of the model. The sigmoid output of the final layer represents the probability of the input being one. The model used Adam optimizer with a learning rate of 0.001 and a binary cross entropy loss function. The implemented ANN model had 427,521 trainable parameters

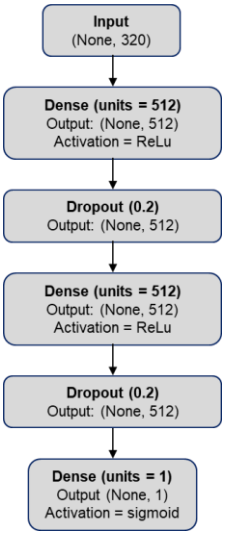

Figure 22: Fully Connected Artificial Neural Network Architecture Diagram

The implemented TCN model architecture is shown in Figure 23. The final TCN model was inspired by Bai et al.'s architecture [39] and was created through an iteration of experiments in varied kernel size, dilations, optimizers, learning rates and stacks while monitoring training accuracy. Early stopping monitoring training accuracy was implemented to allow the TCN to train until 100% training accuracy was achieved. The model with the simplest architecture and lowest training time meeting the aforementioned goal was implemented for the final model. The final TCN model had a kernel size of 6,

residual blocks with dilations of 1, 2, 4 and 8 filters in each convolution layer. A stack for this architecture consists of four residual blocks with the specified dilations. Overall, the model contained one stack of residual blocks yielding a model with 119,425 parameters. The rectified linear unit (ReLU) activation function was used after each dilated convolutional layer followed by channel normalization and spatial dropout of 0.05. The final layer is a fully connected layer with one unit and a sigmoid activation function. Given the problem was binary classification, binary cross entropy was utilized as the loss function. Finally, the model has an Adam optimizer with a learning rate of 0.002 and clip norm of 1.

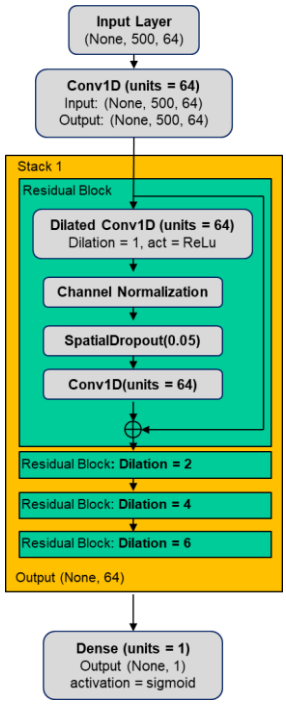

Figure 23: Temporal Convolutional Network architecture diagram.

The implemented LSTM architecture is shown in Figure 24 and was inspired by Popov and Fomenkov's architecture [40]. The LSTM model was twice the size of the

TCN with 213,761 trainable parameters. The first hidden layer of the LSTM model was identical to the TCN by using a Conv1D with a kernel size of 10, dilation of 1, and 64 filters. The final dense layer was also identical with one unit and a sigmoid activation function. The LSTM model consisted of one LSTM layer with 64 units, dropout of 0.2 and recurrent dropout of 0.2. The LSTM layer does not return a sequence, which allows the model to return the one value rather than a sequence. The LSTM model also had the same Adam optimizer with a learning rate of 0.001 and the same binary cross entropy loss function.

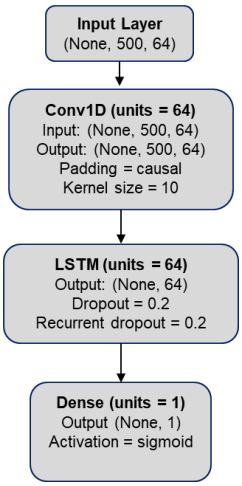

Figure 24: LSTM architecture diagram

3.4.3.2 Training Analysis

To create a model with high performance on a binary classification task, it is necessary to include observations in the training split with the target variable present (i.e. a label of 1) as well as observations with the absence of the target variable (i.e. label of 0). Including both forms of the target variable allows the model to learn important features during the presence and absence of the target variable. One of the main

challenges with the task datasets is the unbalanced distribution of data. As shown in Table 1, the experiment contains three types of the Stand decision-task, four types of the Comparison decision-task, three of the Intel, and four of the HR. Ideally, the training, validation, and test sets should each have the presence and absence of the target variable for each form of the decision-tasks present in the dataset. With the distribution of tasks, it is not possible to have such a distribution of the target variable in a training, validation and test set without separating observations from the same tasks. In addition, the distribution of classification for every task varies for each participant. For example, if a participant was biased on two of the three "Stand" decision-tasks it is not possible to train and test on observations with the presence and absence of the target variable. Thus, any test performance for such a distribution of data is inadequate in reporting the true performance of a model.

Another challenge with the information selection datasets is the small quantity of observations. Depending on the participant, the number of observations range from 40 to 90. Such a small samples size makes test results less significant. Using a test split of ten percent yields only four to eight test observations and if the model classifies one of these observations incorrectly, the performance of the model drastically changes. Also of concern with such a small dataset is the potential of overtraining. These considerations lead to following training and performance evaluation approach.

To compensate for the aforementioned limitation cross-validation was utilized to report model performance. In the task datasets, a leave one task out cross validation (LOTOCV) approach was implemented. In LOTOCV, observations from an entire task were left out of the training set and used as a validation observation. This resulted in 14-

fold cross-validation. The models were fit with the training set, and validation performance was evaluated with observations from the one task left out. This process was repeated until every task had been used as the validation task once. For the information selection datasets, a standard 10-fold cross-validation was applied. Unless otherwise specified, the reported performance throughout this work is the mean validation performance of all iterations of the cross-validation process.

A downside to using cross-validation to evaluate performance of the models is the model cannot be both tuned and evaluated using the validation set. Using the validation set to tune the models would cause information leakage and result in inflated performance metrics during evaluation. Although not ideal, to create suitable models, the batch size and model structure were tuned by observing the training accuracy. Early stopping with the criteria of 100% training accuracy was utilized for model training to ensure models could learn the dataset.

3.4.3.2.1 Cross-Participant

Within-participant machine learning results are usually better than cross-participant results because there is less variability in a single participant's data than across multiple subjects. The benefit of creating cross-participant models is that there is more data to train models and a high performing model is more robust because it generalizes well across multiple participants. To determine if cross-participant models can benefit from the increased data with hyper-parameter optimization, cross-participant models will be explored on the highest performing within-participant dataset.

For the cross-participant training, a train, validation and test dataset approach was applied. In this approach, 12 participants were used as a train set, two participants were

used as a validation set for hyper-parameter optimization and finally the model performance was measured on the remaining participants sequestered data. This process was repeated 15 times so that every single participant's data was the test set. The same hyper-parameters explored for within-participant models were tuned for cross-participant. The major difference being the hyper-parameters were optimized using the validation dataset accuracy rather than the training accuracy. This process allowed for cross-participant test performance to be reported rather than cross-validation performance as was reported for within-participant models.

3.4.3.3 Performance Analysis

To measure the ML model's performance, classification accuracy is observed. Accuracy provides a measure of how well each model performs across both classes. With the class imbalance another beneficial performance metric is balanced accuracy**.** Balanced accuracy accounts for the class imbalance and is the average recall obtained on each class. This metric provides better insight into model performance on both classes. To provide more clarity on the type of classification errors, confusion matrices will be employed (Figure 25). Confusion matrices allow the true positive rate (TPR), true negative rate (TNR), false positive rate (FPR) and false negative rates (FNR) to be visually observed which provides insight into what type of classification errors the model is making.

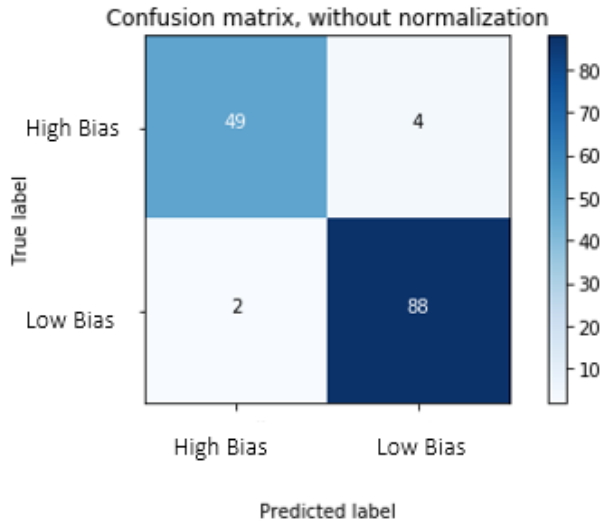

Figure 25: Notional Confusion Matrix for Model Performance on Bias Detection

Since accuracy and confusion matrices are dependent upon tuning the threshold used to classify a model's probability output, the measure which will be used to determine the overall best model is the Area Under the Receiver Operating Characteristic Curve (AUROC) as shown in Figure 26. A ROC curve is created by plotting a model's TPR vs. FPR at many possible classification thresholds. An AUROC of 1.0 indicates a model of high performance. AUROC will be used to determine how well the relationships in the data for classifying bias or information selection from EEG signals is modeled.

In addition to performance metrics another important aspect of performance analysis is error analysis. For the information selection datasets, the main focus for error analysis will be on analyzing how classification errors are associated with a specific task. If confirming and disconfirming information from specific tasks consistently produce classification errors, then models may not be generalizing across tasks in the assessment. For the task datasets, the main focus of error analysis will be on data segmentation. There are numerous ways to segment epochs from the entire task. For future work, it is

necessary to determine if observations should only be taken from the information search portion of a task or the entire task.

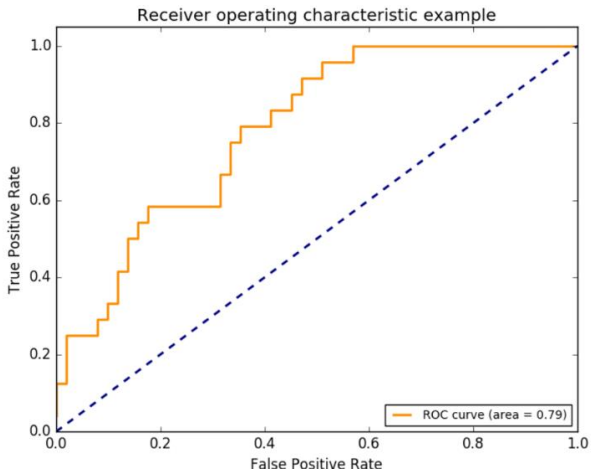

Figure 26: Notional Receiver Operating Characteristics Curve Summary

3.5 Summary

In summary, this chapter outlined the methodology for a human-subject decision-making experiment and established a ML pipeline to estimate confirmation bias or information selection. The decision-making experiment consisted of participants completing a modified ABC test, which elicits biased behavior in nine decision tasks. The test was modified to only elicit confirmation bias as well as to time synchronize the collected physiological data. Analysis methods to quantify the degree of confirmation bias in each task was established. In addition, methods to model relationships between behaviors and biased decisions were designed. Finally, the complete machine learning pipeline was devised which covered everything from raw data preprocessing and feature extraction to model design and performance analysis.

IV. Analysis and Results

4.1 Chapter Overview

This chapter provides in-depth analysis of the results obtained from the experiment outlined in Chapter 3. Results include behavioral measures of confirmation bias collected during the ABC assessment. Section 4.2 describes the subjective results of the assessment, which includes information search patterns, the effects of an initial decision on information search, evidence and question importance, completion time, and information revisits. These results provide support for answering investigative questions one and two (see Section 3.2). Section 4.3 describes non-subjective measures, covering results associated with the electrophysiological data collected during the assessment. Results in this section describe the machine learning performance metrics of classifying electroencephalography (EEG) signals as well as the more traditional EEG time series analysis. These results provide justification in answering investigative questions three and four (see Section 3.2).

4.2 Behavioral Analysis and Results

Behavioral analysis entails all non-physiological components recorded by the ABC assessment. Behaviors of interest include:

- Information Search
- Evidence/Question Importance
- Completion Time
- Uncertainty

Information search and evidence/question importance are the measures used to quantify the degree of bias present in each respective task. In tasks, which require information search, the effect of making an initial decision prior to the information search is analyzed to determine if the desired effect of a more balanced information search was achieved. In addition, the cross-participant mean level of bias for each type of task is analyzed to determine if different types of task are more practical to create high and low levels of bias. It is important to note the levels of confirmation bias quantified in these results are not directly comparable to the results reported by MITRE [42]. MITRE reports an overall bias score for each of the different elicited biases, whereas this work reports only a confirmation bias score for each task.

Multiple aspects of completion time are of interest including completion time of the information search portion of a decision task, and time to complete the entire assessment. These aspects of completion time are analyzed to determine if they are associated with the degree of bias in a task and across tasks. Lastly, uncertainty is measured by mouse re-clicks and could indicate a participant's indecisiveness by a participant's tendency to select different information multiple times. All the described behavioral measures are investigated for being correlated with biased information search in the following sections.

4.2.1 Information Search

In the ABC assessment, the Stand and Comparison form of tasks had the participant perform information search to complete the task. With three versions of the Stand task and four versions of the Comparison task, this yielded seven tasks where the

participant was instructed to search through information. Information for these tasks was composed of confirming and disconfirming information relative to their initial decision. For tasks without an initial decision, the information was confirming and disconfirming relative to their final decision. Table 8 shows the mean and standard deviation of the proportion of confirming information selected during the information search portions of the Stand tasks and the Comparison task for each participant. Figure 27 shows the cross-participant mean proportion of confirming information for the Stand and Comparison tasks with 95% confidence intervals. The Stand tasks overall cross-participant mean proportion of confirming information selected is 0.52 (standard deviation of 0.25) and the Comparison tasks is 0.65 (standard deviation of 0.24). These results indicate that on average during the information search portion of these decision-tasks, participants selected more confirming information than disconfirming information.

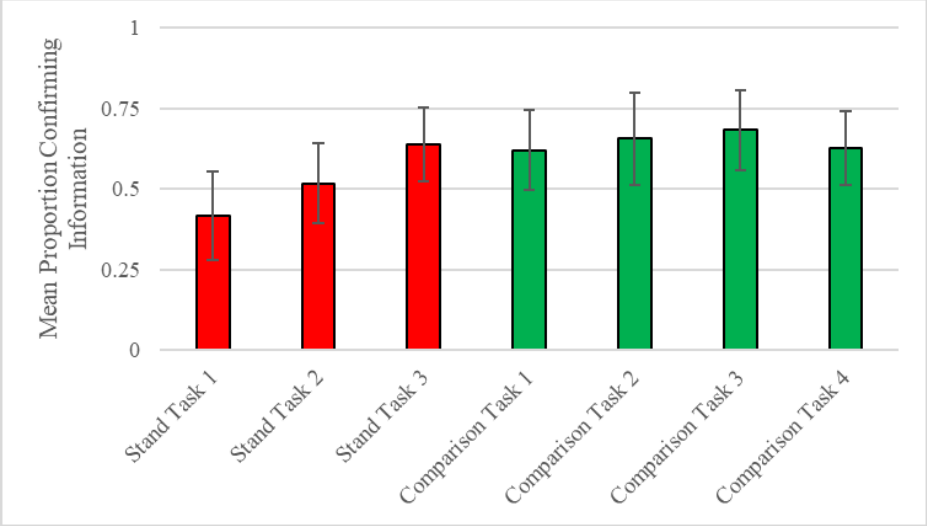

Figure 27: Cross-participant Information Selection

From a machine learning perspective, there are several results of interest in the information search patterns. In particular, in the Stand tasks participant 3097 had a mean proportion of 0.50 with a standard deviation of 0.00. Using a threshold of greater than 0.50 to classify a task as biased would yield all three of the Stand tasks as unbiased. Similarly, for the same subject the Comparison task mean proportion is 1.00 with a standard deviation of 0.00 yielding all four comparison tasks as biased. A machine learning classifier with the goal of classifying a task as biased will not likely perform well on this participant because every observation from the Stand form of tasks is identical and every observation from the Comparison form of tasks are identical. Ideally, each form of task should have a similar number of biased and unbiased tasks.

Table 8: Information Proportion in Information Search Tasks

Participant ID	Stand Task: Confirming Info Proportion			Comparison Task: Confirming Info Proportion		
	Samples	Mean	Stand Dev	Samples	Mean	Stand Dev
1234	17	0.56	0.08	36	0.68	0.12
1962	8	0.22	0.16	12	0.47	0.17
3097	20	0.50	0.00	10	1.00	0.00
3914	15	0.60	0.14	48	0.67	0.00
4818	16	0.36	0.05	40	0.65	0.03
4960	13	0.64	0.10	23	0.65	0.09
6809	24	0.50	0.00	30	0.72	0.11
6910	7	0.72	0.21	19	0.75	0.18
6920	20	0.51	0.07	28	0.77	0.08
7344	11	0.44	0.08	10	0.56	0.41
7590	12	0.83	0.24	21	0.50	0.13
7914	20	0.58	0.12	48	0.67	0.00
7958	4	0.33	0.47	7	0.67	0.50
7960	10	0.39	0.28	24	0.49	0.30
9646	18	0.67	0.24	23	0.58	0.10
Cross-Participant	215	0.52	0.25	379	0.65	0.24

4.2.1.1 Initial Decision Effect on Information Search

In an attempt to create a balanced distribution of biased and unbiased tasks, the experiment manipulated the independent variable of the presence of an initial decision prior to information search; see Table 6 for an outline of the independent variable. In the Stand form tasks, one task did not have an initial decision while two tasks had an initial decision. With only three total tasks, statistical tests were not feasible and thus the observed trends are observational. Across all participants, the mean proportion of selected confirming information in Stand tasks with an initial decision is 0.47 (standard deviation of 0.24), while the mean confirming proportion for Stand tasks without an initial decision is 0.64 (standard deviation of 0.22). Figure 28 illustrates these means with 95% confidence intervals.

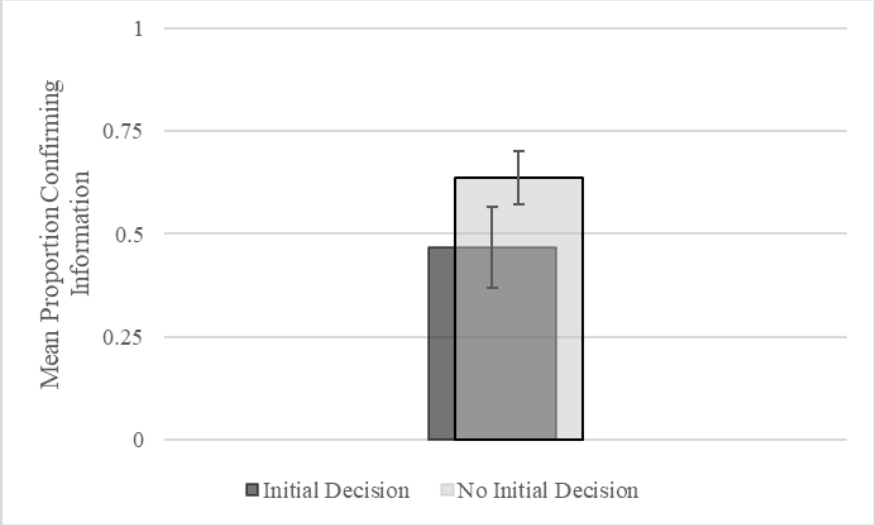

Figure 28: Stand Task Proportion Confirming Information

These results suggest when participants made an initial decision prior to conducting information search, the participants selected less confirming information compared to tasks without an initial decision. These findings were not the results that

were hypothesized in research question one. It was expected that making an initial decision would result in a more balanced information search because information search is conducted post-decision which is when rationalization of the decision occurs and seeking of more information may be biased toward the previous decision (see Section 2.2). A possible explanation for the more balanced information search post-decision is equal amounts of confirming and disconfirming information were present which may have caused participants to question their decision. Cognitive dissonance occurs in the post-decision phase of the decision-making process and occurs when one has inconsistent thoughts about their decision and as a result believes an alternative decision was better. With no choice being favored over the other based on the amount of information present, participants may have experienced cognitive dissonance and thus explored more disconfirming information.

4.2.2 Evidence/Question Importance

In the ABC assessment, two forms of tasks did not contain an information search step in the decision task. These tasks elicited biased behavior by having participants select evidence or questions they believed were most important in completing the task. The forms of tasks that elicited bias by evidence and question importance were the Intel and the Human Resources (HR) tasks respectively. In these tasks, the participant's goal is to answer questions and the tasks contained evidence/questions either confirming or disconfirming with respect to the question. With three Intel tasks and four HR tasks, the assessment contained seven tasks that elicited biased behavior by the participant's value of evidence/questions.

The degree of confirmation bias in a task is the proportion of confirming evidence/questions selected by the participant. Table 9 shows the mean and standard deviation of the proportion of confirming evidence/questions for all participants in the Intel and HR tasks. Figure 29 shows the cross-participant mean proportion of confirming evidence/questions for the Intel and HR tasks with 95% confidence intervals. The Intel tasks' cross-participant mean proportion of confirming evidence in the Intel task is 0.71 (standard deviation of 0.17) and the HR tasks' is 0.59 (standard deviation of 0.16). The results indicate participants tended to select more evidence that confirmed the questions asked in the Intel tasks, but only slightly more confirming questions in the HR tasks.

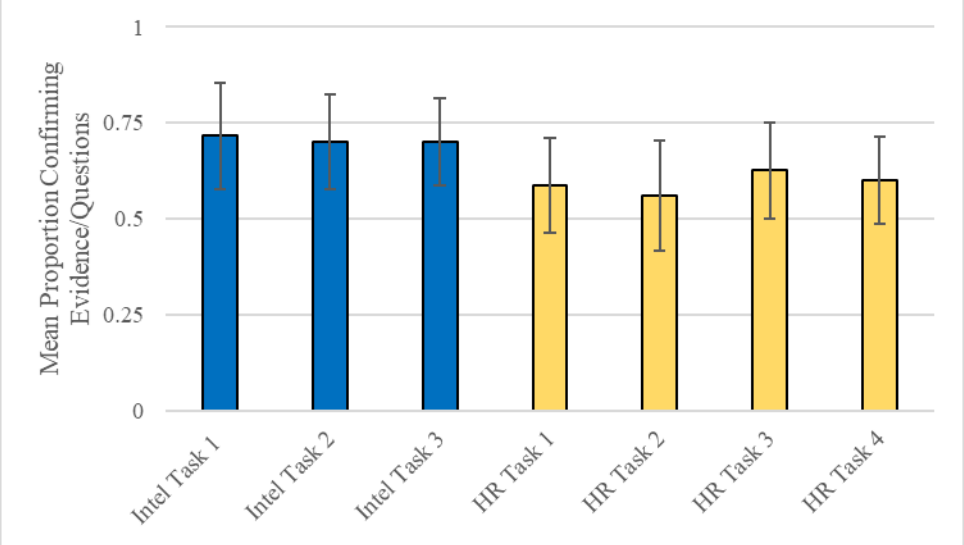

Figure 29: Cross-participant Evidence/Question Importance

Table 9: Information Proportion in Evidence/Question Importance Tasks

Participant ID	Intel Tasks: Confirming Info Proportion			HR Tasks: Confirming Info Proportion		
	Samples	Mean	Stand Dev	Samples	Mean	Stand Dev
1234	4	0.92	0.12	5	0.50	0.17
1962	4	0.58	0.12	5	0.70	0.10
3097	4	0.92	0.12	5	0.60	0.00
3914	4	0.83	0.12	5	0.50	0.22
4818	4	0.67	0.12	5	0.50	0.10
4960	4	0.67	0.12	5	0.50	0.10
6809	4	0.58	0.12	5	0.70	0.10
6910	4	0.58	0.12	5	0.40	0.14
6920	4	0.67	0.12	5	0.75	0.09
7344	4	0.83	0.12	5	0.60	0.14
7590	4	0.58	0.12	5	0.60	0.14
7914	4	0.75	0.00	5	0.70	0.10
7958	4	0.83	0.12	5	0.65	0.17
7960	4	0.50	0.00	5	0.65	0.17
9646	4	0.67	0.12	5	0.55	0.17
Cross-Participant	60	0.71	0.17	75	0.59	0.16

An interesting observation of the data in Table 9 is the much lower standard deviation for the evidence/question importance tasks (Intel and HR) compared to the information search tasks (Stand and Comparison). The lower standard deviation for the Intel and HR task is likely due to the set number of evidence/questions selected by the participant. The participant could only select four/five pieces of evidences/questions respectively. On the contrary, in the Stand and Comparison tasks the participant had to select at least one piece of information but was free to select as many pieces of information that were available. The smaller standard deviation for the evidence/question importance tasks indicate a machine learning bias classifier may not perform well on these tasks given the data is unbalanced.

4.2.2.1 Provided Hypothesis Effect on Evidence Importance

In an attempt to create a more balanced distribution of biased and unbiased questions, the independent variable for the Intel task was the presence or absence of an accepted hypothesis. The HR tasks did not have any control variable manipulation in the assessment due to the tasks lack of a hypothesis or decision primer. In each Intel task, the participant selected important evidence to answer a question. Along with the evidence, an accepted hypothesis was provided to give the participant a specific stance on the subject. In two of the three Intel tasks, the accepted hypothesis was absent. The mean proportion of confirming evidence selected for Intel tasks without an accepted hypothesis is 0.71 (standard deviation of 0.17) and with an accepted hypothesis is 0.70 (standard deviation of 0.16). Figure 30 illustrates these means with 95% confidence intervals.

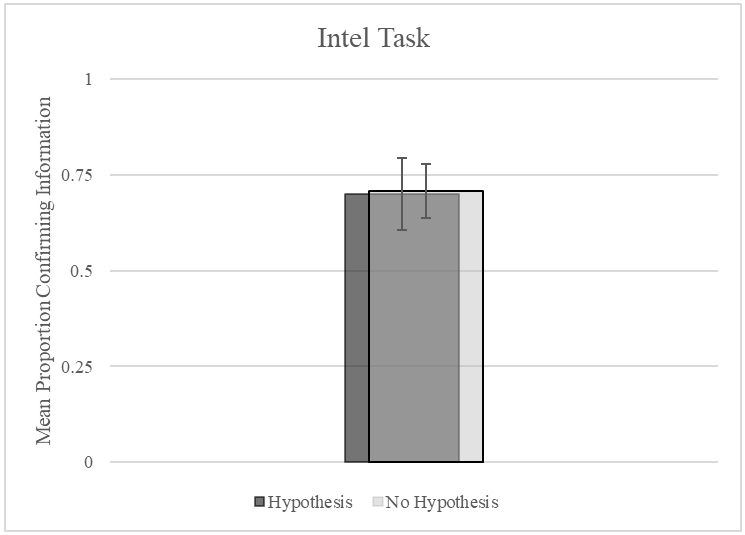

Figure 30: Intel Task Proportion Confirming Evidence

The lack of difference of means between the two groups suggests the intended effects did not occur for the Intel tasks. Both groups had almost identical measures despite the lack of hypothesis in one group. A possible cause of the lack of disparity

between the groups could be due to the rigid constraint in how many pieces of evidence the participant could choose. Participants could have only wanted to choose less than four pieces of evidence, but the constraint could have resulted in more evidence selected and thus different confirming proportions. Regardless, the only conclusion that can be drawn from these results is the absence of a hypothesis compared to the presences of a hypothesis did not affect a participant's value of evidence in the Intel tasks.

4.2.3 Completion Time

Analysis of participant behavioral data was completed to explore all collected data. The subjective information search and evidence/questions importance results in the previous sections provide a means to measure level of confirmation bias, but analysis of the behavioral data may provide further insight into behavioral characteristics associated with confirmation bias. The first aspect of completion time explored was the cross-participant completion time for each of the 14 tasks. Identifying any trends in the completion time across participants could indicate if completion time is a valuable behavior in characterizing biased behavior. Figure 31 shows the cross-participant mean completion time for each task in the assessment.

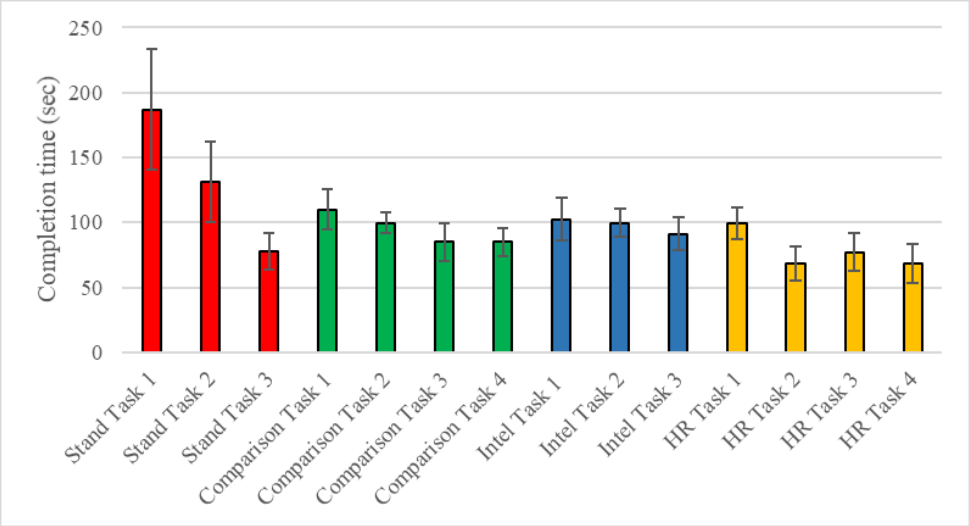

Figure 31: Cross-participant Task Mean Completion Time

The task label along the horizontal axis corresponds to a tasks relative order in the assessment i.e. Stand Task 1 appeared before Stand Task 2 and Stand Task 3, but not necessarily before Comparison Task 1, see Table 6 for the overall order of tasks in the assessment.

Apparent in Figure 31 the mean completion time for each type of task decreases with the order of the presented task. A one-way ANOVA was applied to each task of the same type to determine if the mean completion time was statistically different. The Stand ($F = 10.49$ $p = 0.002$), Comparison ($F = 3.55$ $p = 0.02$), and HR ($F = 4.46$ $p = 0.007$) tasks all had statistical significance while the Intel task ($F = 0.71$ $p = 0.49$) did not. To analyze if the mean task completion time decreased with respect to the order of the presented task, a left tail paired *t*-test was utilized. The first input into the *t*-test was the set of participant completion times on Stand task 1 and the second input was the paired set of participant completion times on Stand task 3. This effectively tests if the participant completion time on Stand task 3 was smaller than Stan task 1. The difference of the

means for Stand task 1 and Stand task 3 is 109.36 seconds and is statistically significant (t = -5.48, p = 0.00004) with a significance level of 0.05. The difference of the means for Comparison task 1 and Comparison task 4 is 25.13 seconds and is statistically significant (t = -2.75 p = 0.007) with a significance level of 0.05. Similarly, the HR task with a mean difference of 31.21 seconds between the first and last HR task were statistically significant (t = -4.38, p = 0.003). The only task without a decrease in completion time with respect to task order was the Intel task which had a mean difference of 11.14 seconds between the first and last task and was not statistically significant (t=-1.62, p = 0.06).

A possible explanation for the observed trend of decreasing task completion time with the order of the task is a learning effect occurred. As participants became more familiar with a task, they were able to learn the task and complete the task more quickly. Another possible explanation for the observed decreasing completion time is the inherent completion time is different for each task. The information associated with each task was not identical in length which may have resulted in different completion time. To rule out this as a possible cause, future work should alternate the task order for each participant. Regardless of what caused the trend in completion time, the observed trend is a confounding effect and is the dominating effect on completion time. For completion time to be associated with confirmation bias, completion times should be relatively the same with respect to order which would allow analysis on the effects of bias on completion time. Due to these confounding effects, models were not created to associate completion time with bias.

The results displayed in Figure 31 encompass the entire decision task, which includes the scenario instructions, initial decision, information search, and final decision. If the observed trend is due to a learning effect, the first type of each task could result in a longer completion time because the participant is reading the scenario instructions whereas in subsequent tasks reading instructions may not be necessary. In an attempt to reduce this effect on completion time, the completion time for only the information search and final decision portion of the task were analyzed. Figure 32 shows completion time for the information search portion of the Stand tasks and Comparison tasks. The Intel and HR tasks were not included because there was not an information search portion of the tasks.

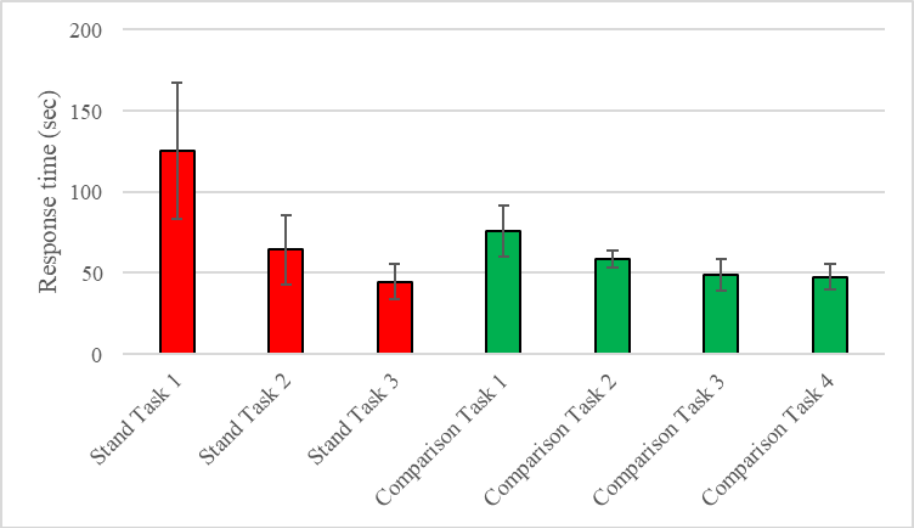

Figure 32: Cross-participant Information Search Completion Time

As shown in Figure 32 the decreasing trend in completion time is still apparent as the order of task increases. The difference in mean completion time between tasks is reduced compared to the entire task, but the decrease trend still appears in information search

completion time. With the apparent confounding variables in completion time, further analysis was focused on other behavioral data.

4.2.4 Information Revisits

Participants utilized the computer mouse to select information in the information search tasks and to record their decisions. All mouse clicks used to select a piece of information or choose an answer were recorded by the ABC assessment platform. Information revisits, or selecting information multiple times prior to making a decision could be indicative of the participants increased uncertainty. Information revisits for this analysis is the cumulative number of times the participant reselects a piece of information and is calculated for a given task with the following formula:

$$Information\ Revisits = Total\ Mouse\ Clicks - Selected\ Information$$

As a data exploration step, the mean cross-participant information revisits on each task was presented in a histogram with 95% confidence interval using a normal distribution as shown in Figure 33.

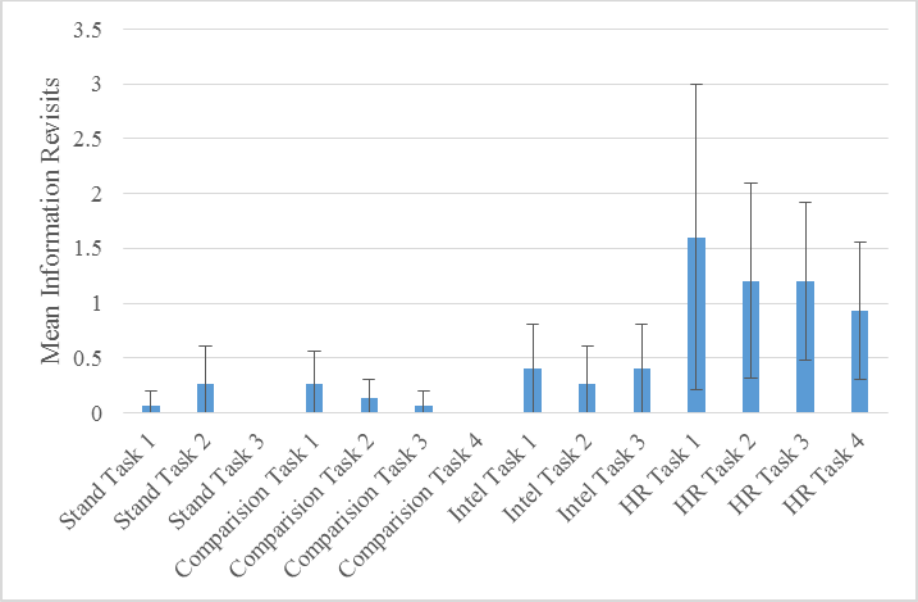

Figure 33: Cross-Participant Information Revisits for each Task

At the cross-participant level, information revisits appear to be a rather low value and does not appear to be useful. The only apparent trend in the data is that the HR tasks had a higher mean information revisits, but the high variability shown by the 95% confidence intervals indicates this was not consistent across participants.

To analyze if there is an association with the participant's uncertainty and proportion of confirming information or evidence/questions selected at the participant level, the correlation coefficient between a participant's information revisits and confirming proportion was measured. Figure 34 illustrates the confirming proportion of selected information plotted against the information revisits for participant 7958.

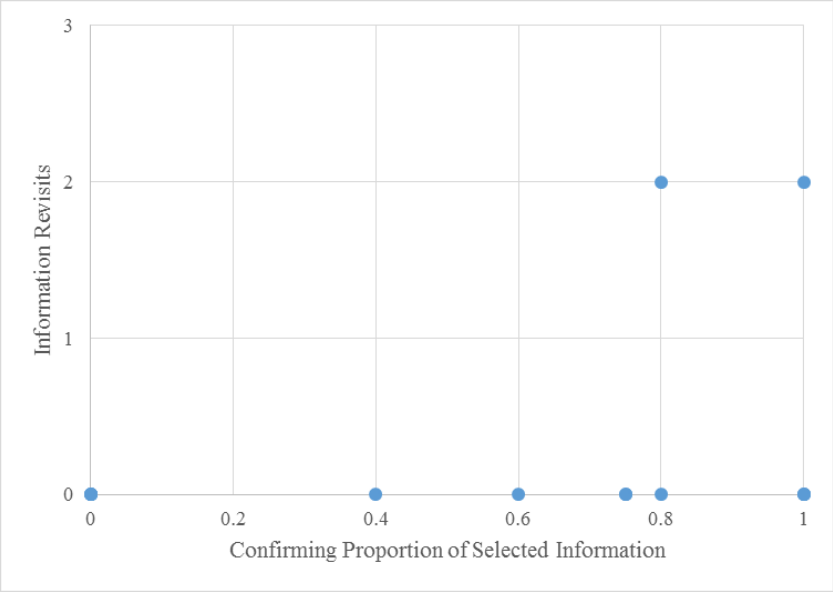

Figure 34: Confirming Information vs. Information Revisits

Of the 15 participants, only two showed correlation between information revisits and proportion of confirming information. Participant 1234 and participant 7958 had a correlation coefficient of -0.52 and 0.42 respectively. Although these coefficients appear to indicate a relationship between excessive information revisits and proportion of confirming information, a majority of the tasks had zero information revisits. Participant 1234 only had information revisits on three tasks and participant 7958 only had information revisits on two tasks as shown in Figure 34. The mean number of information revisits for all participants across the Stand, Intel and HR tasks is 0.63 with a standard deviation of 1.38. The lack of information revisits across all tasks indicate the behavioral response of information revisits is not suitable for association with bias in the respective decision tasks.

4.3 Electroencephalography Analysis and Results

4.3.1 Machine Learning: Task Dataset

In the early data exploration stages, there were apparent patterns in the Task datasets that were not ideal for achieving high performance with machine learning methods. The most prominent pattern in this dataset is the imbalance of decision-tasks labeled as biased. The most drastic imbalance being 89% of the observations biased for participant 7914 and the least being 39% biased for participants 1962. Table 10 displays the number of observations and the distribution of biased observations for each participant.

Table 10: Task Datasets Class Distribution

Participant ID	Observations	Positive Class (%)
1234	642	0.694
1962	570	0.397
3097	427	0.763
3914	457	0.722
4818	585	0.553
4960	528	0.733
6809	811	0.488
6910	295	0.463
6920	444	0.754
7344	472	0.712
7590	521	0.510
7914	370	0.890
7958	403	0.564
7960	379	0.411
9646	612	0.448

Due to the imbalance of data, the machine learning results for this dataset are expected to be highly dependent on the participant and their specific distribution of data. For

example, participant 3097's Comparison tasks are all classified as biased meaning when any comparison task is used to measure validation accuracy it will likely be near 100%. Whereas in participant 1234, only Comparison task 1 is unbiased so when Comparison task 1 is used as the validation task the accuracy will be near 0% because the models were trained with only biased Comparison tasks.

4.3.1.1 Time Series Features

The Task dataset with time series features is the Time Series Signal per Task dataset referred to in Section 3.4.2. For this dataset, an observation is a 2-second time series signal (500 frames at 250 Hz) segmented from the information search portion of a task. There are 64 features, which are the 64 EEG electrode locations positioned on the participants head. The labels for this dataset is "bias" or "unbiased". The total number of observations vary by participant, ranging from 295 to 811 observations. The results reported in this section are cross-validation performance metrics. The cross-validation implemented is a leave one task out cross-validation resulting in 14 validation folds.

The TCN and LSTM models implemented for the Time Series Signal per Task dataset were trained using early stopping, based on 100% training accuracy, for a maximum of 10 epochs. On average, the TCN models trained for 7.3 epochs and the LSTM models trained for 6.5 epochs. With the TCN average time per epoch being 2 seconds and each participant having 14 models trained, the average TCN training time per participant was 204 seconds. The LSTM average time per epoch was 6 seconds, which yielded an average training time of 546 seconds per participant. Despite the

training time difference between the models, both the TCN and LSTM were able to obtain 100% accuracy on the training data as displayed in Figure 35.

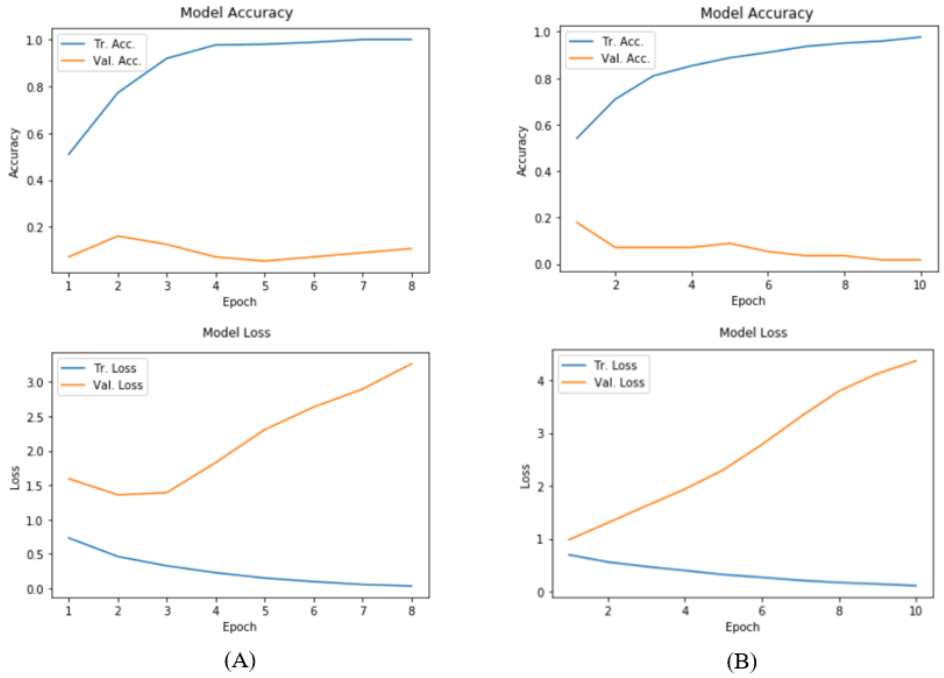

Figure 35: Participant 7958 (A) TCN Training Curves (B) LSTM Training Curves

The left images (labeled A) of Figure 35 correspond to the TCN model while the right images (labeled B) correspond to the LSTM model. Both models for this specific participant obtain similar validation accuracies of approximately 10%, indicating both models are unable to correctly classify the validation data correctly.

The poor performance illustrated in the training curve is also reflected in the overall accuracy of the models. The accuracy plot in Figure 36 shows the mean leave one task out cross-validation classification accuracy for each participant. The red bar represents the majority class percentage or the baseline accuracy if all decision-tasks were classified as the majority class. The green and purple bars are the TCN and LSTM

overall accuracies. The error bars on the TCN and LSTM bars represent the 95% confidence interval using a normal distribution.

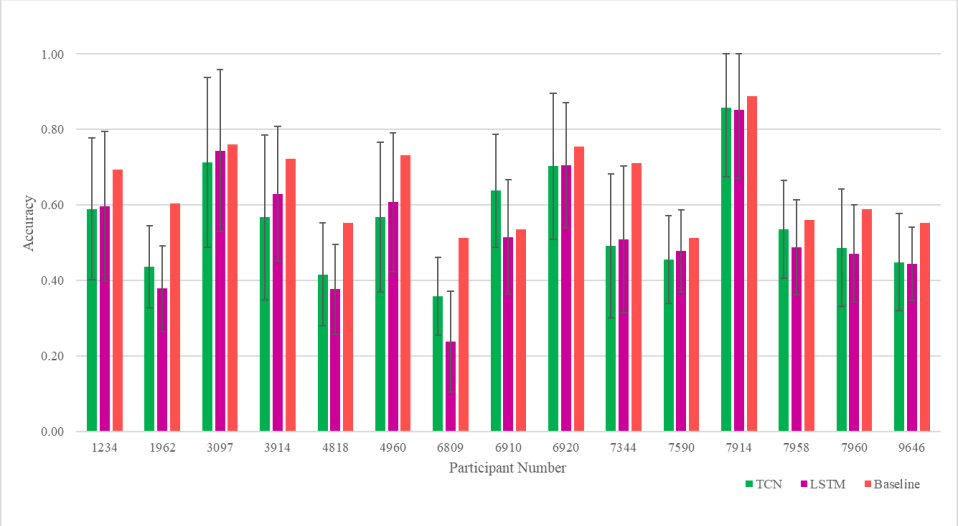

Figure 36: Time Series Signal per Task Model Accuracy

Among 15 participants, the TCN model only achieved a mean accuracy greater than the baseline accuracy on participant 6910, but the TCN mean accuracy of 0.634 ± 0.150 on this participant was not statistically significant (95% Normal Confidence Interval). The LSTM model did not achieve a mean accuracy greater than baseline accuracy on any participants. These results suggest it is not possible to estimate the presence of confirmation bias on this task using the EEG data segmentation and feature techniques in the Time Series Signal per Task dataset.

In a dataset with a class imbalance, overall accuracy can be a skewed metric for evaluating model performance. Balanced accuracy accounts for the class imbalance and is the average recall obtained on each class. This metric provides better insight into model performance on both classes and is necessary since both biased and unbiased task estimation is necessary. The reported balanced accuracy on the Task datasets is the

balanced accuracy of all model predictions from each cross-validation step rather than the mean balanced accuracy at each cross-validation step. Since the validation fold only consists of observations from the same task all validation observations have the same label. Reporting the mean balanced accuracy would be the same metric as overall accuracy. The balanced accuracy displayed in Figure 37, shows models were able to obtain a balanced accuracy greater than 50% on two participants: 6910 and 6920. On participant 6910, the TCN obtained a balanced accuracy of 0.674 and the LSTM balanced accuracy was 0.501. For participant 6920, only the LSTM obtained balanced accuracy greater than 50% with an accuracy of 0.552.

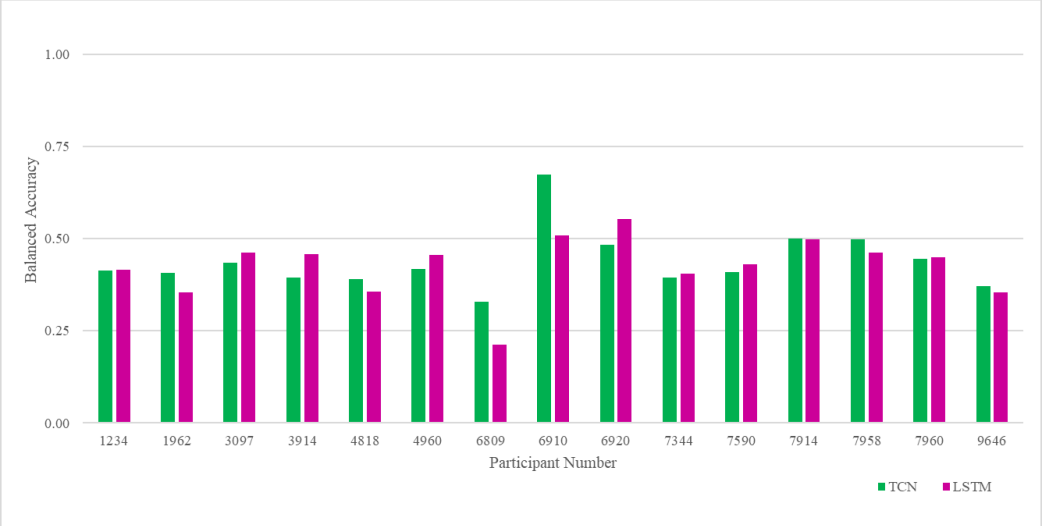

Figure 37: Time Series Signal per Task Model Balanced Accuracy

Accuracy and balanced accuracy represent the possible performance metrics at specific thresholds. The area under the receiver operating characteristic curve (AUROC) of a model shows how well a model learns the relationships in the dataset. Figure 38 displays the AUROC for the TCN and LSTM for each participant.

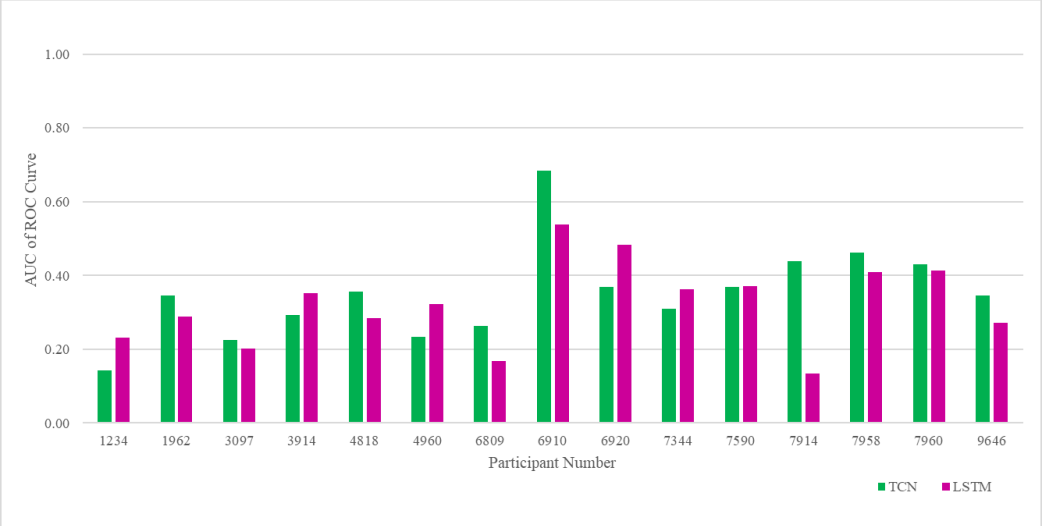

Figure 38: Time Series Signal per Task Model AUROC

The average AUROC across all participants for the TCN and LSTM models were 0.351 and 0.322 indicating the models had low performance on the dataset.

Observing the confusion matrices (Figure 39) for the top two performing participants in balanced accuracy illustrates the impact of unbalanced data on the model performance. Participant 6910's data was 46% biased and is reflected in the results by both models predominantly predicting "Not Biased". Participant 6920's data was 75% biased and is reflected in both models predominantly predicting biased. The poor model AUROC and lack of consistent model performance across all participants indicates either the time series of the EEG data or the method of segmenting the data by task with the label of biased or unbiased is not an appropriate method for measuring confirmation bias.

Since observations for this dataset are from the entire information search portion of a decision task, the fundamental assumption is there is a consistent difference in brain activity between biased and unbiased during the entire decision-making process. One theoretical explanation of cognitive biases which directly counters this assumption is the

neuroscientific perspective. The neuroscientific perspective relates the occurrence of various cognitive biases to principles that are characteristic of biological neural networks and consequently are a result of the neural characteristics of the brain [13]. This perspective conjects that cognitive biases occur in the same neural networks as motor functions and thus there is no distinguishable brain activity relating to a biased decision. While the lack of high performance for this dataset does not indicate the neuroscientific perspective on cognitive biases is true, if there is no distinctive brain activity directly from biases, estimating confirmation bias on this dataset would not work.

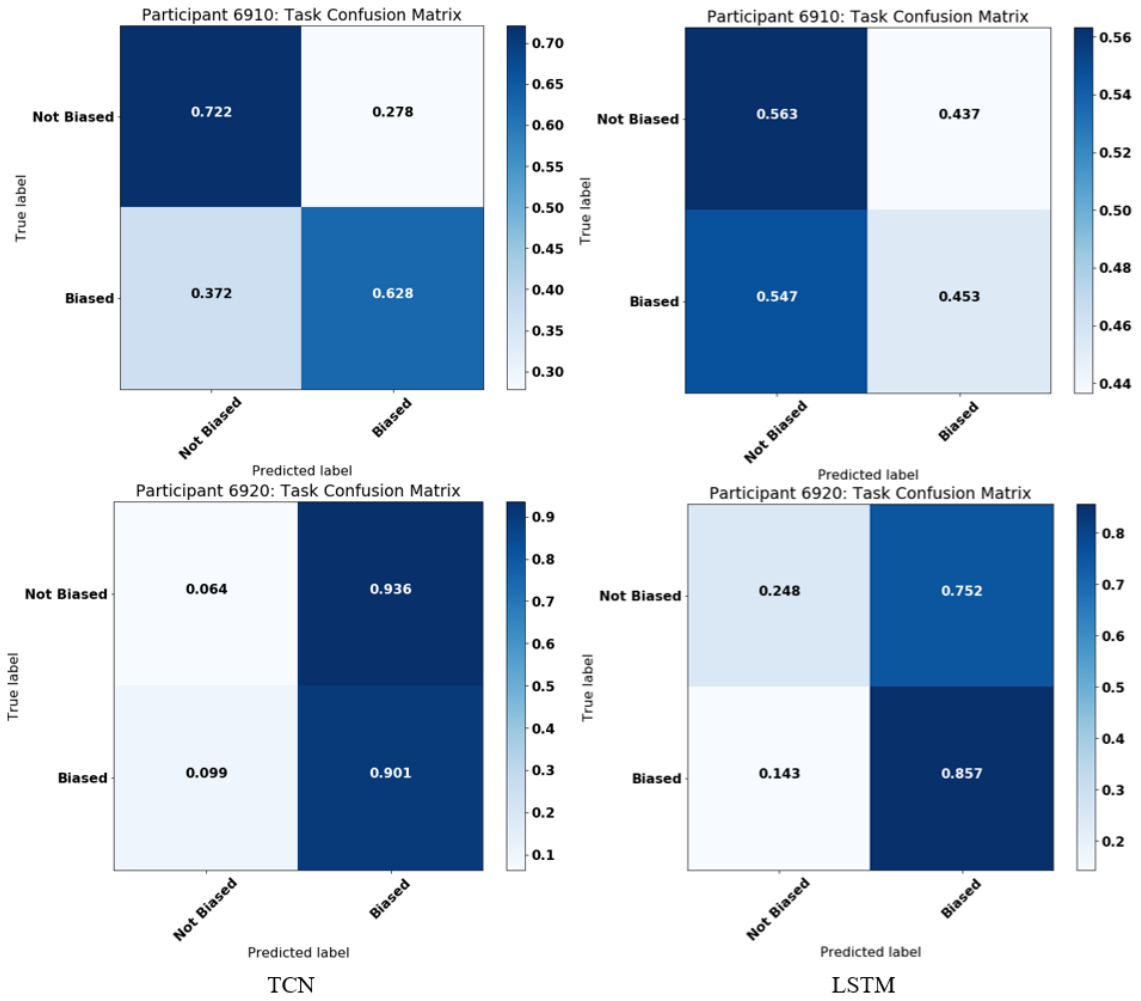

Figure 39: Confusion Matrices of (Left) TCN, (Right) LSTM

4.3.1.2 Mean Frequency Power Features

The mean frequency power features on the task dataset is Frequency Features per Task dataset referred to in Section 3.4.2. In this dataset, the features are the mean power of the five frequency bands at each electrode in the 64 electrode EEG cap, totaling 320 features. The frequency feature extraction (described in Section 3.4.1.2) was implemented on the 2-second time series signal from the Time Series Signal per Task dataset in the previous section. For this reason, the observations, data distribution, and

leave one task out cross-validation were identical to the Time Series Signal per Task dataset in the previous section.

Given the mean frequency power features were extracted from the same two-second time series in the previous section, results were expected to be similar but possibly better due to the specific frequency extraction. The ANN trained for a maximum of 10 epochs with early stopping based on training accuracy resulting in an average of 3.5 epochs. With 14 models trained per participant and an average epoch time of 1 second, the average ANN training time was 49 seconds per participant. The overall accuracy for the ANN, LDA, and RFC models is shown in Figure 2Figure 40. All models performed similarly across participants with the mean accuracy of LDA being 0.540, RFC 0.532 and the ANN 0.531. LDA performed the best across participants with accuracy above baseline on participants 3097, 4960, 6920. Unfortunately, none of these accuracies were significantly greater than baseline (95% Normal Confidence Interval). RFC obtained better than baseline accuracy on participants on 3097 and 6920 but were also not significantly greater than baseline (95% Normal Confidence Interval). Lastly, the ANN only performed better on participant 6920 and was not significantly greater than baseline (95% Normal Confidence Interval). Although the overall accuracy for any participant was not statistically significant, the mean frequency power features appear to perform marginally better on the task dataset than the time series feature.

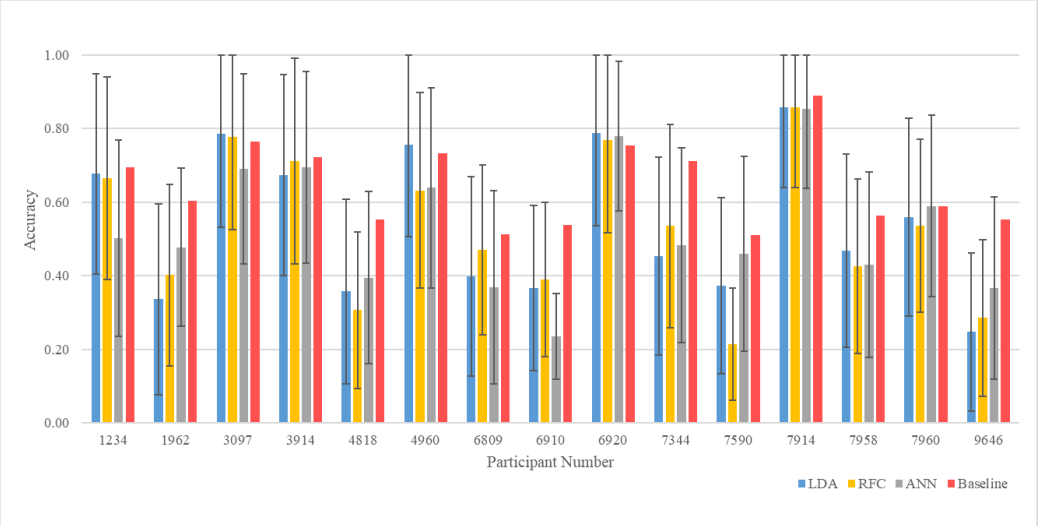

Figure 40: Frequency Features per Task Model Accuracy

The improved performance of the frequency feature over the time series feature is also reflected in the balanced accuracy metric (Figure 41) with at least one model achieving greater than 50% balanced accuracy on participants 4960, 6920, and 7960. The ANN model had the highest balanced accuracy of 0.596 on participant 6920, which was also the highest performing participant in the time series feature for the LSTM model. Both the LDA and ANN performed above 50% accuracy on participant 7960, which was not a high performing participant for the time series feature. The change in performance across features indicates the time series feature may not be optimal for classifying bias from EEG signals.

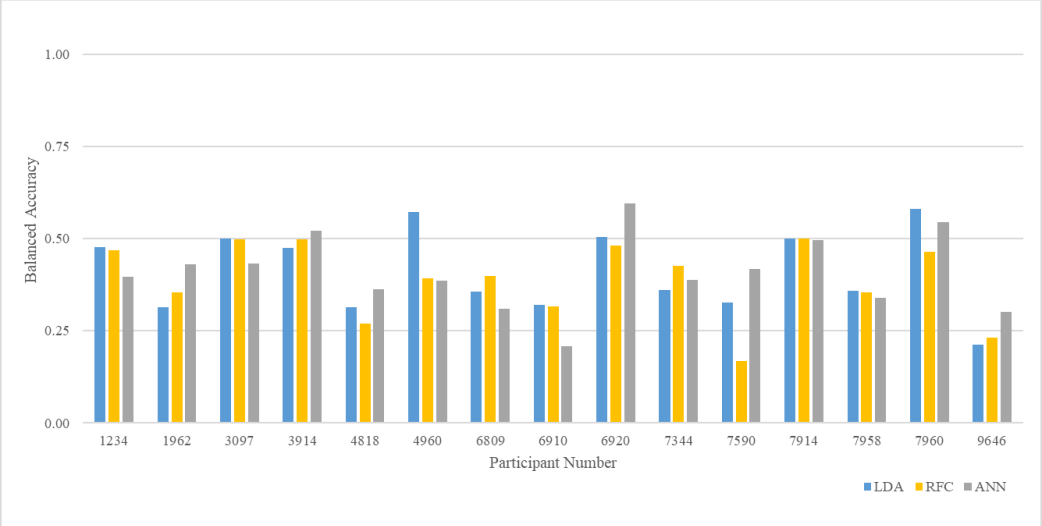

Figure 41: Frequency Features per Task Model Balanced Accuracy

The average AUROC across all participants for LDA, RFC, and the ANN were 0.356, 0.259, and 0.340 respectively. Although the AUROC for each model indicates low model performance as displayed in Figure 42, the ANN achieved an AUROC of 0.785 for participant 6920. This outlier in performance is likely due to the highly imbalanced data.

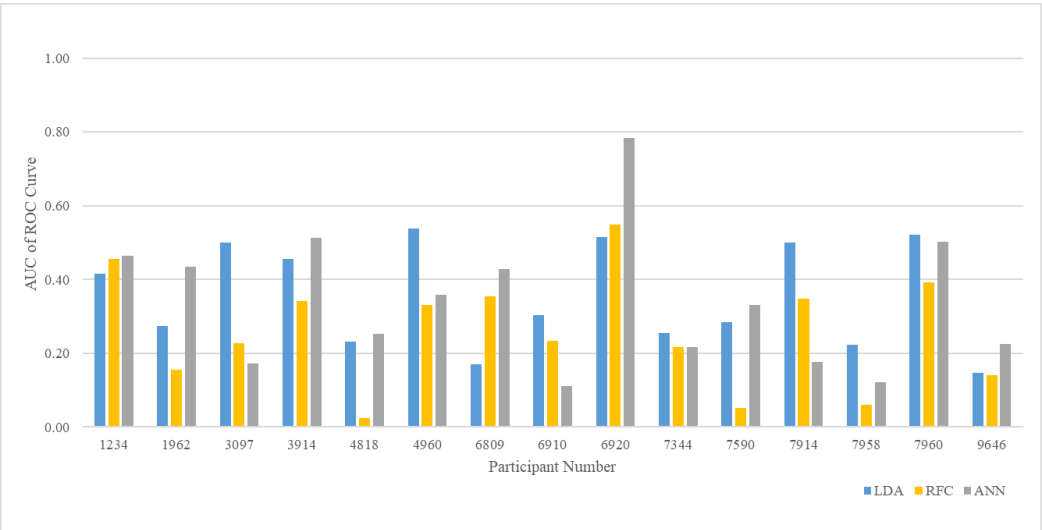

Figure 42: Frequency Features per Task Model AUROC

As shown in Figure 43, the confusion matrix for the ANN model on participant 6920 indicates the model is over classifying the “bias” label class. The classification imbalance is due to participant 6920’s dataset being highly imbalanced with 75% labeled “biased”. The high AUROC for participant 6920 indicates with a proper training, validation, and test set, tuning the classification threshold may obtain greater accuracy and balanced accuracy. Despite this performance anomaly, the AUROC across all participants is low. Although the frequency features performed marginally better than the time series features, performance was still under the desired baseline of 50% balanced accuracy. If there is no constant difference in brain activity under biased and unbiased decision during the decision-making process as proposed in the neuroscientific theory of cognitive biases, then the EEG signals under both conditions are likely similar. With no distinctive difference between the EEG signals, the machine learning models would be unable to distinguish between the two labels “biased” and “unbiased” resulting in worse than chance performance. For this reason, even though there were multiple more ways to investigate the Task dataset, the poor performance and limited time constraints led the investigator to focus on the Information dataset in Section 4.3.2.

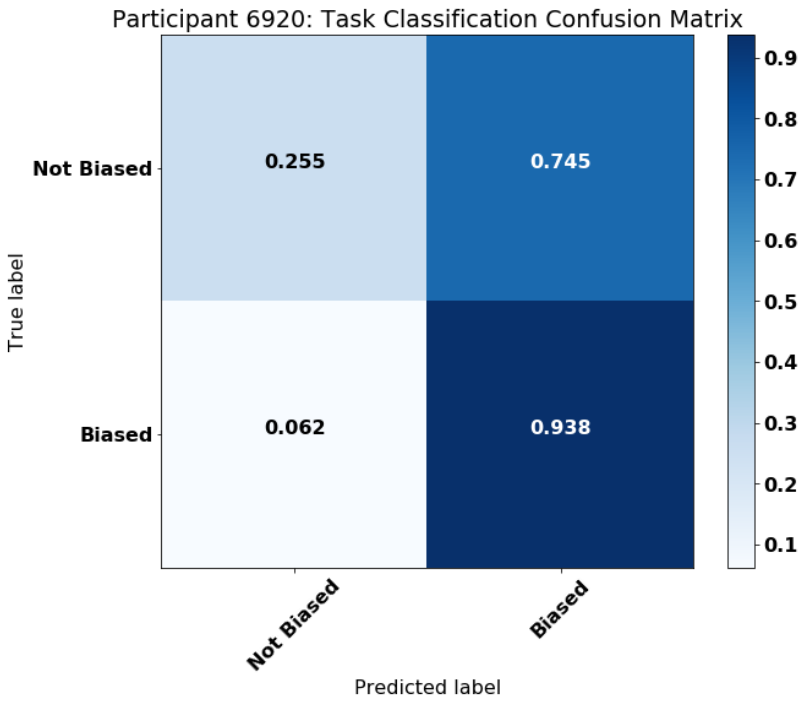

Figure 43: ANN Task Frequency Confusion Matrix

4.3.2 Machine Learning: Information Selection Datasets

The Information datasets consists of EEG signals when the participant selected confirming and disconfirming information. Compared to the Task datasets, the Information Selection datasets were much more balanced as shown in

Table 11. The positive class is the class label "confirm" and is the majority class across all participants. The highest percentage of positive class was 0.746 for participant 3097 while the lowest was 0.557 for participant 7344. The machine learning challenge for this dataset is the number of observations rather than a data imbalance; as illustrated in Table 11 participant 3914 has the most observations with 93 and participant 7958 has the least with 42.

Table 11: Information Selection Datasets Class Distribution

Participant ID	Observations	Positive Class (%)
1234	78	0.615
1962	53	0.585
3097	59	0.746
3914	93	0.677
4818	70	0.586
4960	66	0.621
6809	71	0.606
6910	62	0.597
6920	75	0.680
7344	61	0.557
7590	65	0.569
7914	61	0.672
7958	42	0.643
7960	67	0.582
9646	71	0.563

4.3.2.1 Time Series Features

The information selection dataset with time series features is the Time Series Signal per Information Selection dataset referred to in Section 3.4.2. One observation is a 2-second (500 frames at 250 Hz) time series signal centered on when a participant selects a piece of information. There are 64 features, which are the 64 EEG electrode locations positioned on the participant's head. The labels are either "confirm" or "disconfirm" and are the position of the information relative to the participant's belief in a task. The number of observations and data balance is displayed in Table 11. The ensuing performance metrics are metrics from 10-fold cross-validation performed on within

participant data. The cross-validation was a stratified split so each fold approximately represented the overall data distribution of "confirm" and "disconfirm" labels.

The implemented TCN and LSTM models for Time Series Signal per Information Selection dataset were trained with early stopping monitoring the training accuracy for maximum of 10 epochs. The TCN trained for an average of 4.3 epochs with an average of 1 second per epoch. The LSTM trained for an average of 5.6 epochs with an average of 10 seconds per epoch. With 10-fold cross-validation, the average training time per participant for the TCN was 43 seconds while the LSTM was 560 seconds. As expected with a small dataset, both models easily obtained a 100% accuracy on the training data as displayed in the training curves in Figure 44.

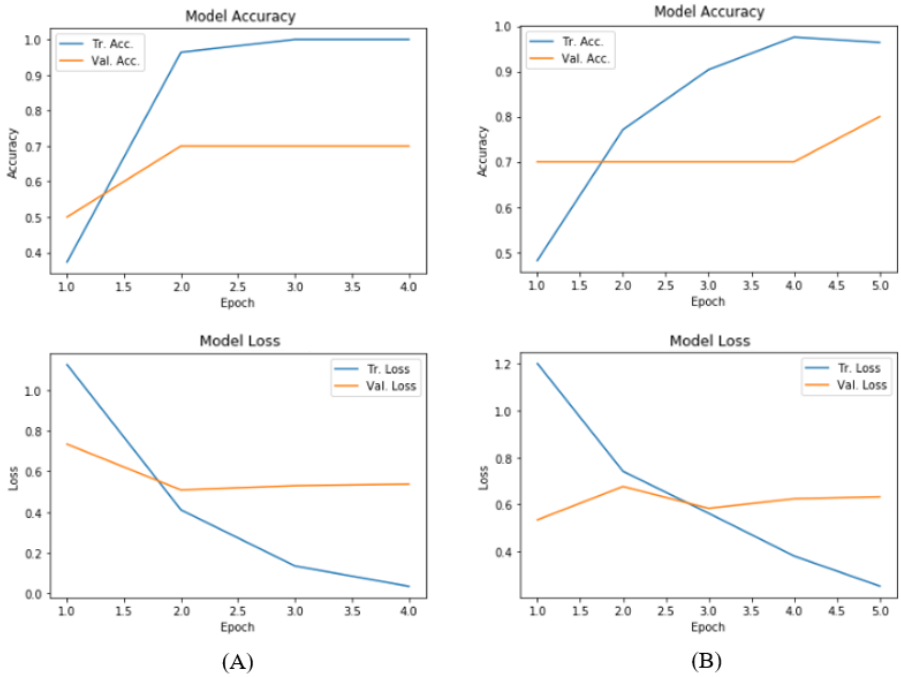

Figure 44: (A) TCN Training Curves (B) LSTM Training Curves

The overall accuracy for the TCN and LSTM models is shown in Figure 45. Of all the 15 participants, the TCN model mean accuracy was only greater than baseline accuracy on participant 6920. The TCN model achieved an accuracy of 0.682, but the difference between baseline was not significant (95% Normal Confidence Interval). One observable difference between the Task and the Information datasets is the much smaller 95% confidence interval error bars for the mean accuracy of the Task dataset. The tighter confidence interval is reflective of the more balanced classes in the Information dataset.

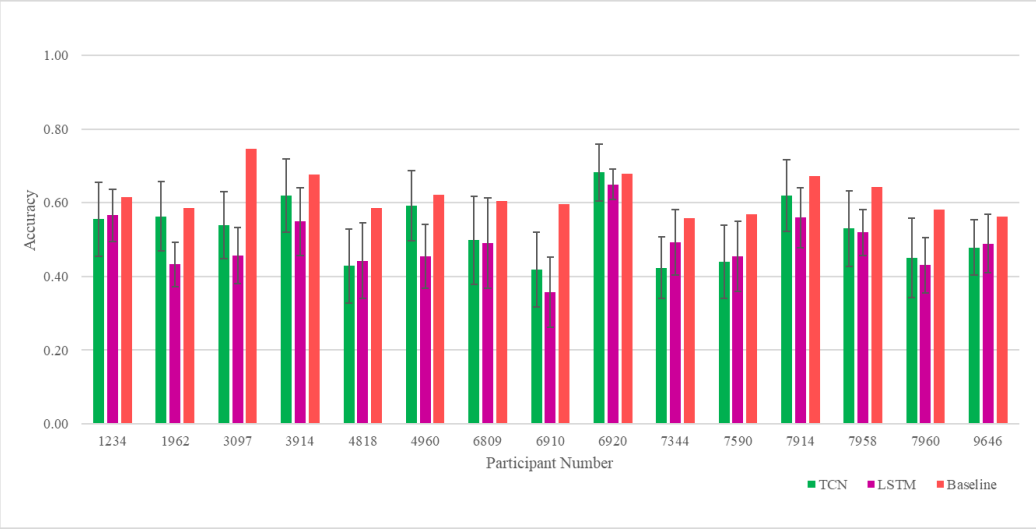

Figure 45: Time Series Information Selection Model Accuracy

Although the TCN and LSTM models were unsuccessful in obtaining higher than baseline accuracy in all but one participant, the balanced accuracy metric was greater than 50% for at least one of the models in seven participants (Figure 46). Considerably greater than the two participants for the Task dataset, the balanced accuracy metric indicates the Information dataset may be more suitable for machine learning. The TCN model appears to perform better than the LSTM across the participants by performing better than 50% balanced accuracy on six participants, while the LSTM only has five

above 50%. Unfortunately, the TCN only performed significantly greater than the 50% balanced accuracy baseline on participant 6920 (95% Normal Confidence Interval). To further interpret the classification error difference between the models the confusion matrices (Figure 47) on participant 1234 shows both models perform identical on the disconfirm class, but the LSTM model mislabels more confirm observations as disconfirm. Overall, model performance observations cannot be made because neither model was tuned, but the TCN appears to perform consistently better at the default settings used for balanced accuracy.

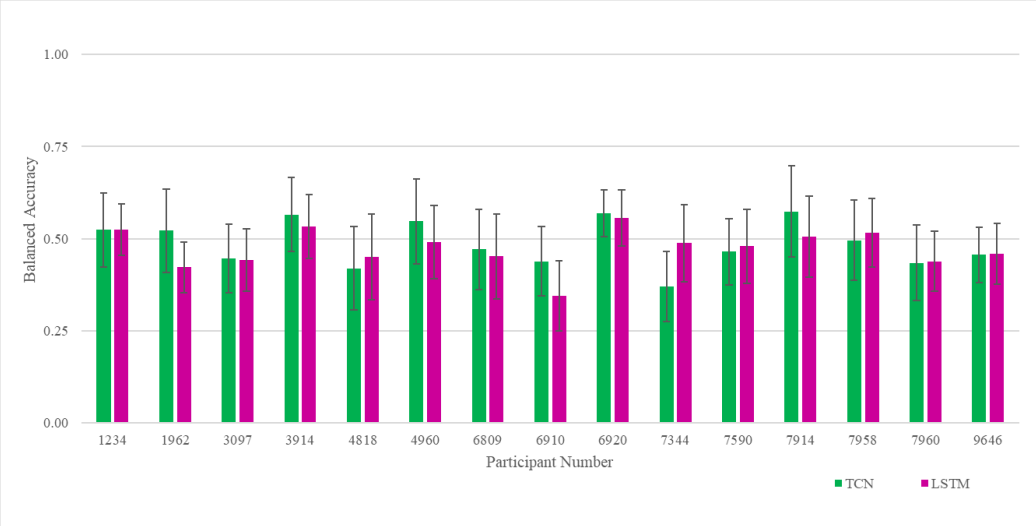

Figure 46: Time Series Information Selection Model Balanced Accuracy

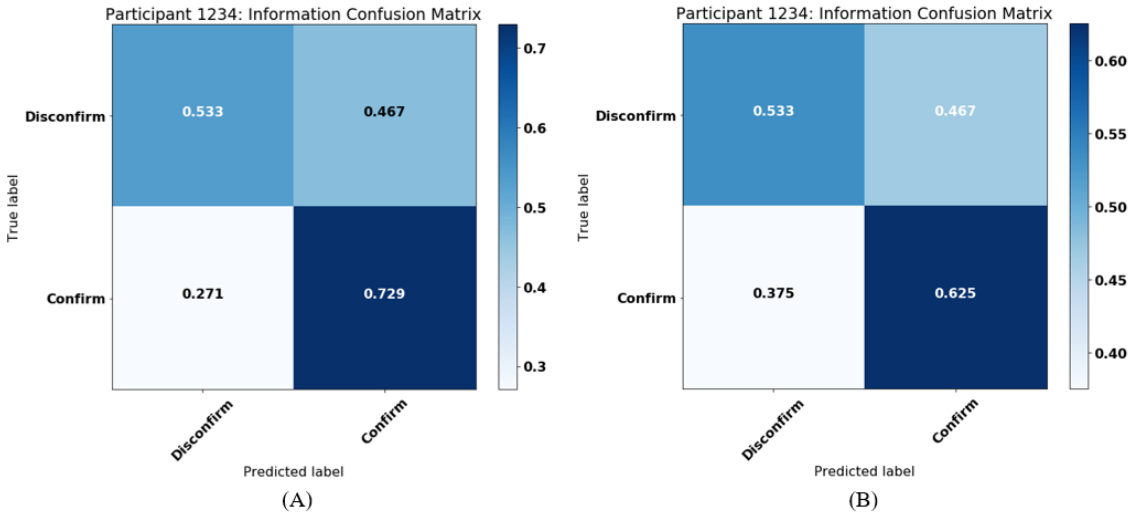

Figure 47: Confusion Matrices (A) TCN, (B) LSTM

Despite the fact that from the balanced accuracy metric the TCN model appeared to perform much better, the mean AUROC across all participants for the TCN was 0.464 and the LSTM was 0.453. As shown in Figure 48, both the TCN and LSTM AUROCs vary per participant.

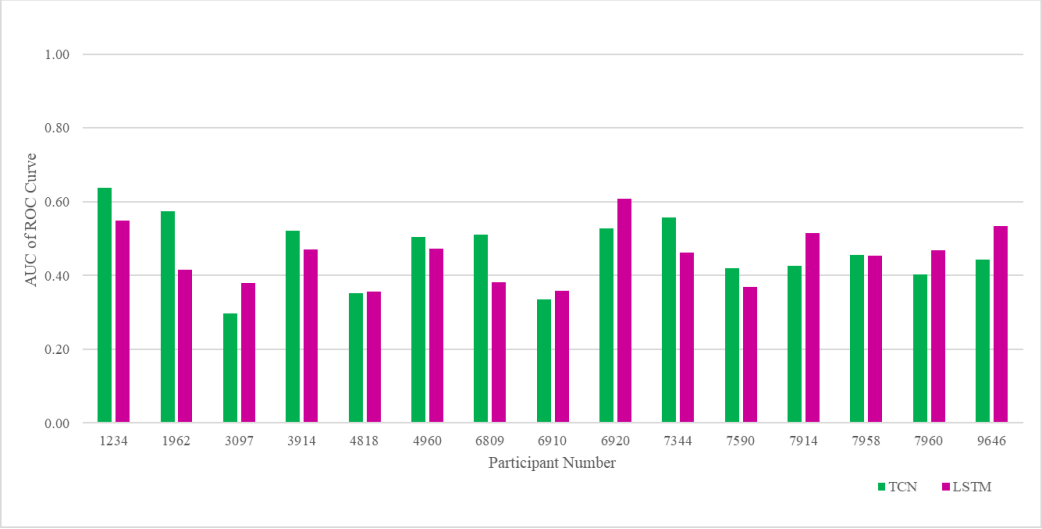

Figure 48: Time Series Information Selection Model AUROC

The increased mean AUROC for both the LSTM and TCN on the Time Series Signal per Information Selection dataset relative to the Time Series Signal per Task dataset indicate the models fit the underlying relationships between the target variable and the features better. Overall, time series classification performance was increased for the Time Series Signal per Information Selection dataset compared to the Time Series Signal per Task dataset, but the still low mean AUROC for both models indicate the time series features do not capture the relationships in the data well.

4.3.2.1.1 Time Series Analysis

To provide further insight into the machine learning performance on the time series features, the cross-participant time series signals were observed. Event-related potentials (ERPs) are commonly used in neuroscience research to compare brain activity between different stimulus. Event-related potentials (ERPs) are voltages created by the brain when experiencing particular stimuli [52]. The time series signals in this work were not necessarily stimulus-locked since the EEG signals correspond to when the participants clicked on a piece of information. For this reason, this analysis refers to the averaged time series signals as time series analysis, but is identical to an ERP except the stimulus lock. As outlined in Section 3.3.6, the difference between cross-participant time series signals from confirming and disconfirming information was tested for statistical significance at the -200 ms to 800 ms time window for the F2, F4, F6 and F8 EEG electrodes. The examined cross-participant time series signals are displayed in Figure 49. In the F4 location there appears to be increased potential for the confirming signal at approximately 275ms relative to disconfirming information. As for the F2 and F8

locations, the presence of disconfirming information appears to have heightened activity compared to confirming information. There does not appear to be any drastic differences between the two types of information at the F6 location. Despite these noticeable differences at the F2, F4 and F8 locations, there were no significant differences between the time series signals of the two information types (nonparametric statistical test using Monte Carlo permutations with cluster corrections, alpha=0.01).

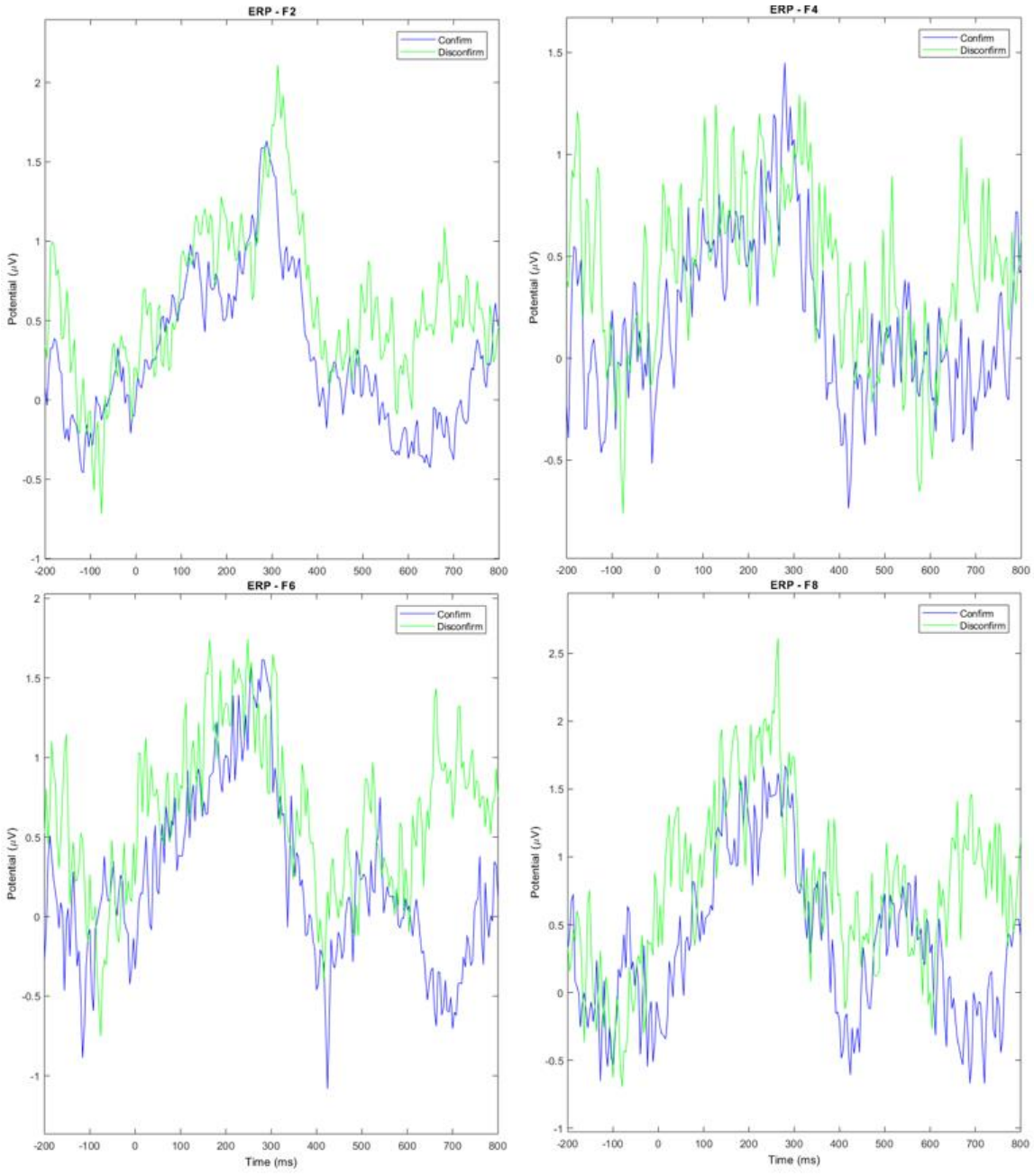

Figure 49: Cross-participant Time Series Signals of Information Type

The lack of difference in the time series signals for confirming and disconfirming information can help explain why the machine learning models performed poorly. Additionally, the cross-participant time series signals are extremely noisy which indicates

there is a significant amount of noise in each participants time series signal. As a data exploration step, the time series signals of the features associated with the right frontal region of the brain were observed for participants which had high and low model performance. Two participant time series signals were observed: participant 6809 and 1962. Participant 6809 had an RFC AUROC of 0.679 while participant 1962 had an RFC AUROC of 0.300 As seen in Figure 50, the relatively higher performing participant had increased activity at the F8 location at approximately 500 ms for confirming information compared to disconfirming information. Whereas for the low performing participant 1962, the time series signal appears much noisier and there are no differentiable trends between the two conditions. These results indicate the poor machine learning results were likely caused by noisy EEG data.

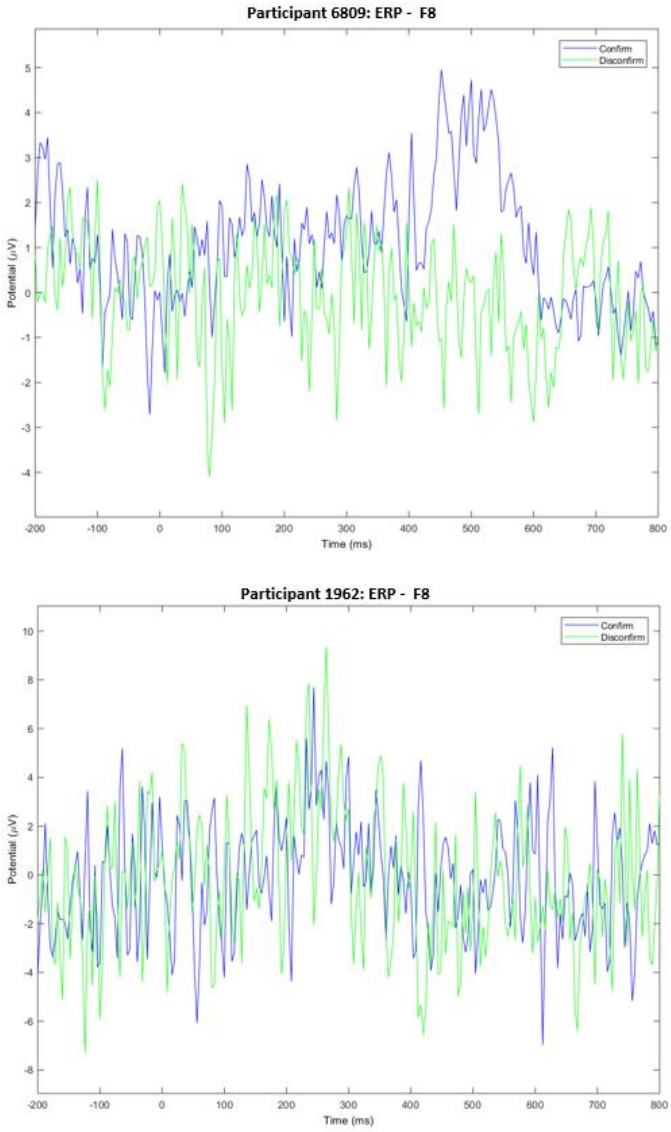

Figure 50: Within-Participant Time Series Signals of Information Type

4.3.2.2 Mean Frequency Power Features

The mean frequency power features on the information selection dataset is Frequency Features per Information Selection dataset referred to in Section 3.4.2. In this dataset, the features are the mean power of the five frequency bands at each electrode in the 64 electrode EEG cap, totaling to 320 features. The frequency feature extraction

(described in Section 3.4.1.2) was implemented on the 2-second time series signal from the Time Series Signal per Information Selection dataset in the previous section. Like the Time Series Signal per Information Selection dataset, a 10-fold stratified cross-validation was implemented to produce the ensuing results.

Due to the increased accuracy of the frequency features over the time series features in the Task dataset, it was hypothesized that the frequency features would outperform the time series features in the Information dataset. The mean accuracy across all participants for the LDA, RFC, and ANN was 0.539, 0.559 and 0.617 respectively. The within-participant mean cross-validation accuracy for each participant is displayed in Figure 51. Overall mean accuracy was greater than baseline accuracy for at least one type of model in seven of the fifteen participants. Although the number of participants with accuracy greater than baseline accuracy was greatly improved from one to seven compared to the time series features, only one participant's mean accuracy was statistically significant. The difference in the ANN mean accuracy of 0.777 ± 0.082 (95% Normal Confidence Interval) and the baseline accuracy of 0.642 on participant 7958 was statistically significant. The considerably smaller confidence intervals and statistical significance in accuracy on one participant indicate the Information dataset with the frequency features may be the highest performing dataset and feature combination explored.

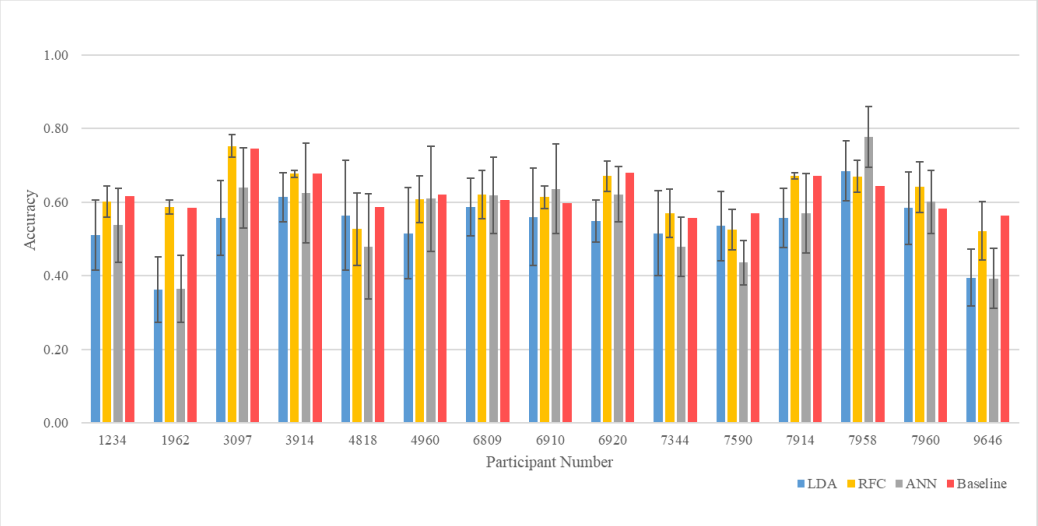

Figure 51: Frequency Information Selection Model Accuracy

The balanced accuracy metric (Figure 52) shows if the model is performing better than random chance of 0.50. At least one model achieved greater than 0.50 balanced accuracy on 10 of the 15 participants. The ANN achieved the highest balanced accuracy of 0.627 on participant 6809, but LDA achieved balanced accuracy greater than 0.50 on nine participants whereas the ANN only achieved greater on seven participants. Despite achieving balanced accuracy above 0.50 on 10 of the 15 participants, the difference between the balanced accuracy and baseline were only significant on two participants. The ANN balanced accuracy was significantly greater than the baseline on participants 6809 and 7960 (95% Normal Confidence Interval). The performance above random chance across multiple participants indicates there is a relationship between the frequency features and the target variable, but the lack of consistent significant difference for the 11 participants indicates more data is likely necessary. It is also important to note, there is a 17% chance of incorrectly obtaining significance on at least two participants (See Appendix A: Balanced Accuracy Incorrect Significance for explanation). In addition, the

balanced accuracy of 62% and 58% on participants 6809 and 7960 are not reliable enough for operational use.

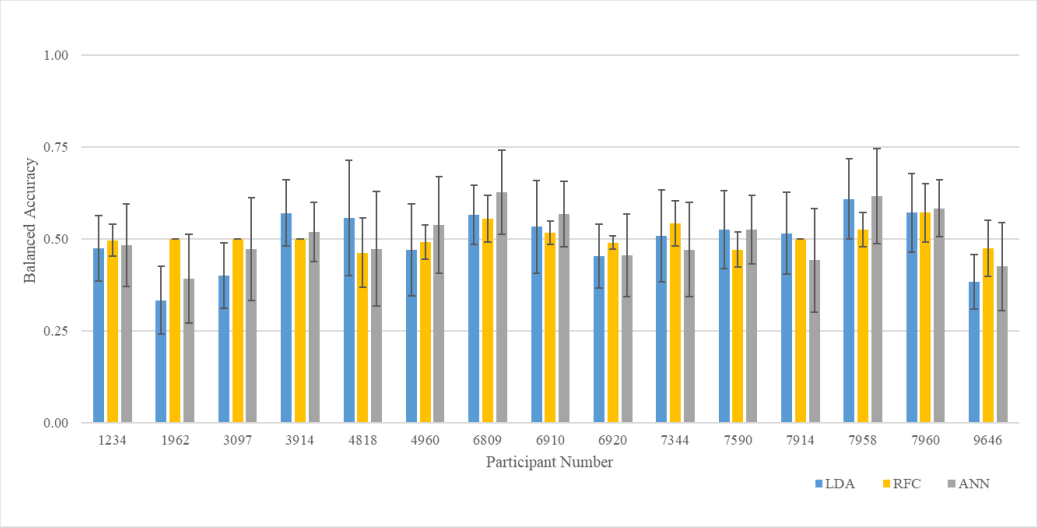

Figure 52: Frequency Information Selection Model Balanced Accuracy

Confusion matrices for the model with the highest balanced accuracy were analyzed to determine how classification errors vary for a model across participants. Figure 53 shows the ANN confusion matrices for participant 6809 and 7958. Participant 6809 has a balanced performance on both disconfirm and confirm class with 53.6% of the disconfirm observations being classified correctly and 67.4% of the confirm observations being classified correctly. Participant 7958 class performance is drastically different with 40.0% of the disconfirm class being classified correctly and 96.3% of the confirm class being classified correctly. The strikingly different class performance can partially be explained by the different class distribution (Table 11). Participant 6809's data was 60.6% of the confirm class while participant 7958's was 64.3%, but the drastic difference is largely due to the observation size difference. Participant 6809 had 71 observations to

participant 7958's mere 42 observations. The effect of the small number of observations is shown by the drastic differences in the model performance across participants.

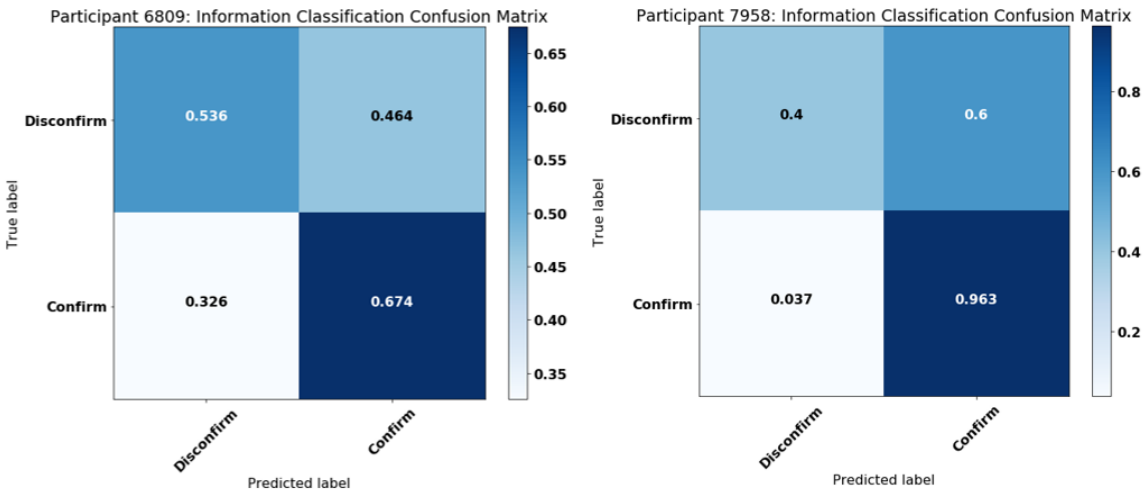

Figure 53: ANN Confusion Matrices

The mean AUROC across all participants for LDA, RFC, and the ANN were 0.493, 0.485, and 0.521 respectively. The AUROC for each model on each participant is displayed in Figure 54. One interesting observation is the participants with the highest AUROC in the frequency feature models, were not the same as the time series models (Figure 48).

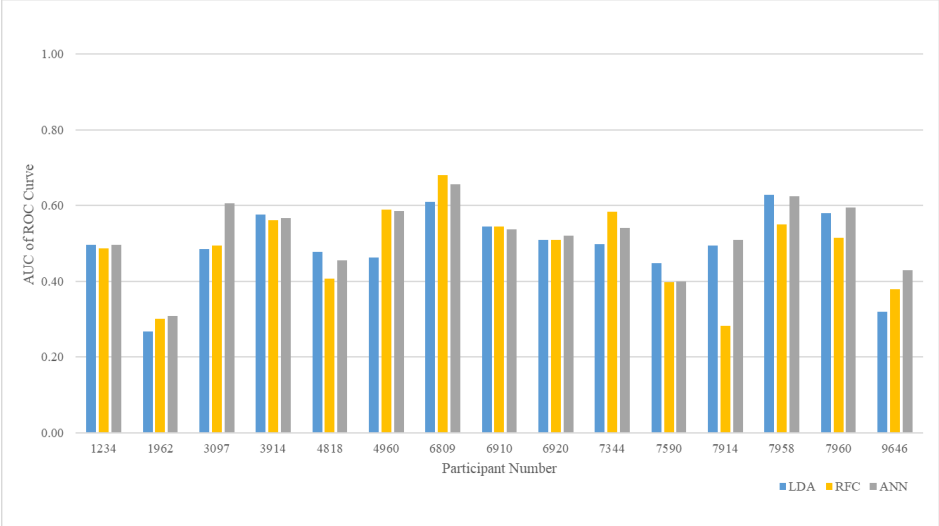

Figure 54: Frequency Information Selection Model AUROC

The mean AUROC for the Frequency Features per Information Selection is much greater than the Frequency Features per Task. The mean AUROC across all participants for LDA, RFC, and the ANN on the Frequency Features per Task dataset (Figure 42) with was 0.356, 0.259, and 0.340 respectively. The increased AUROC for the Frequency Features per Information Selection dataset indicates the method of estimating confirming and disconfirming information to measure confirmation bias is more suitable for machine learning than estimating a biased and unbiased task with the collected EEG data. Regardless of the improved model performance for the Frequency Features per Information Selection dataset, the small sample size for each participant is reflected in the low performance metrics across the participants.

4.3.2.2.1 Feature Importance

Model performance metrics are necessary for determining how well machine learning techniques can model relationships between features and target variables, but feature importance provides crucial details on the specific relationships in the features

being used by the machine learning models. Random Forest Classifier (RFC) is useful because it provides feature importance when fitting the data. Activation of the right frontal portion of the brain that has been associated with the presence confirming information. This region of the brain corresponds to the F2, F4, F6 and F8 features of this dataset. Table 12 shows top ten features based on the number of times a frequency feature appears in a participants 50 most important features. No feature locations correspond to the expected brain location of activity as illustrated in Figure 55.

Table 12: Salient Features across all Participants

Feature Location/Frequency	Count
O1/Delta	7
F1/Theta	7
C2/Delta	7
TP9/Gamma	6
PO8/Delta	6
PO7/Delta	6
O2/Theta	6
C3/Delta	6
C1/Theta	6
FT8/Beta	6

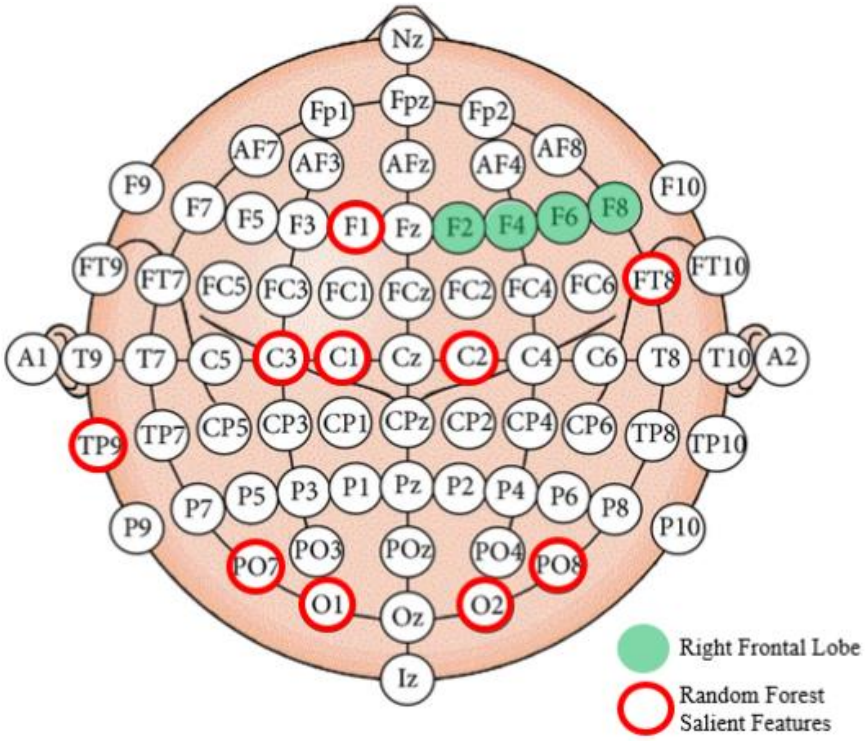

Figure 55: EEG Electrode Locations with Salient Features

To analyze what features were common among RFC which had the highest performance on participants, common features among participants with an RFC AUC greater than 0.50 were analyzed. With this limitation, common features among the nine participants were explored. Table 13 shows the features and the number of times the channel appeared in the top 50 most important features for a participant with the specified threshold of RFC performance.

Table 13: Salient Features in Top performing Participant Models

Feature Location/Frequency	Count
F1/Theta	7
C2/Delta	5
AFz/Alpha	5
TP9/Gamma	4
F8/Beta	4
F6/Theta	4
PO8/Delta	4
O2/Theta	4
C3/Delta	4
C1/Theta	4

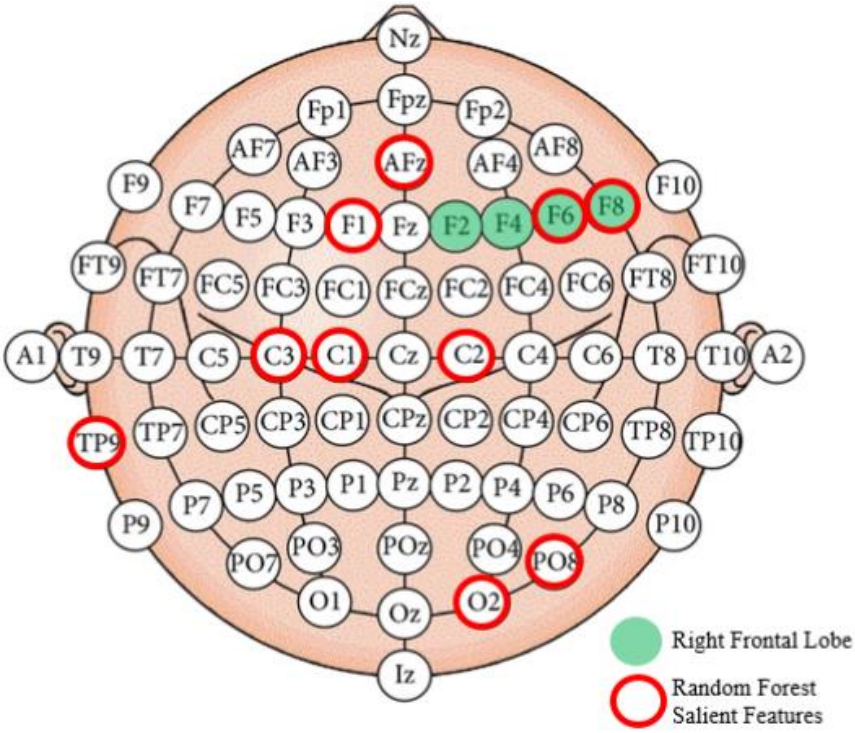

Figure 56: EEG Electrode Locations and Salient Features with High Performance

Two features corresponding to the right frontal area of the brain, F8 and F6, appeared four times in the top 50 features across the nine participants. Not only did one of the F2, F4, F6 and F8 features appear in the top 50 most important features, but at least one of

the expected features were at least one of the top eight most important features in four of the participants with the highest RFC performance (Figure 57). Participants 4960 and 7958 both had the F8 beta frequency as the eighth and fifth most important feature respectively. While participant 6910 had F4 alpha as the sixth most important feature and participant 7344 had F6 theta as the eighth most important feature. The lack of consistent feature importance across all participants indicates either the machine learning techniques were not able to associate expected brain activity with the target variable or the expected brain activity was not consistent across participants. Due to the expected features being present in the eight most important features across four of the participants with the highest RFC performance, the lack of common features across all participants may be due to the small number of observations for each participant.

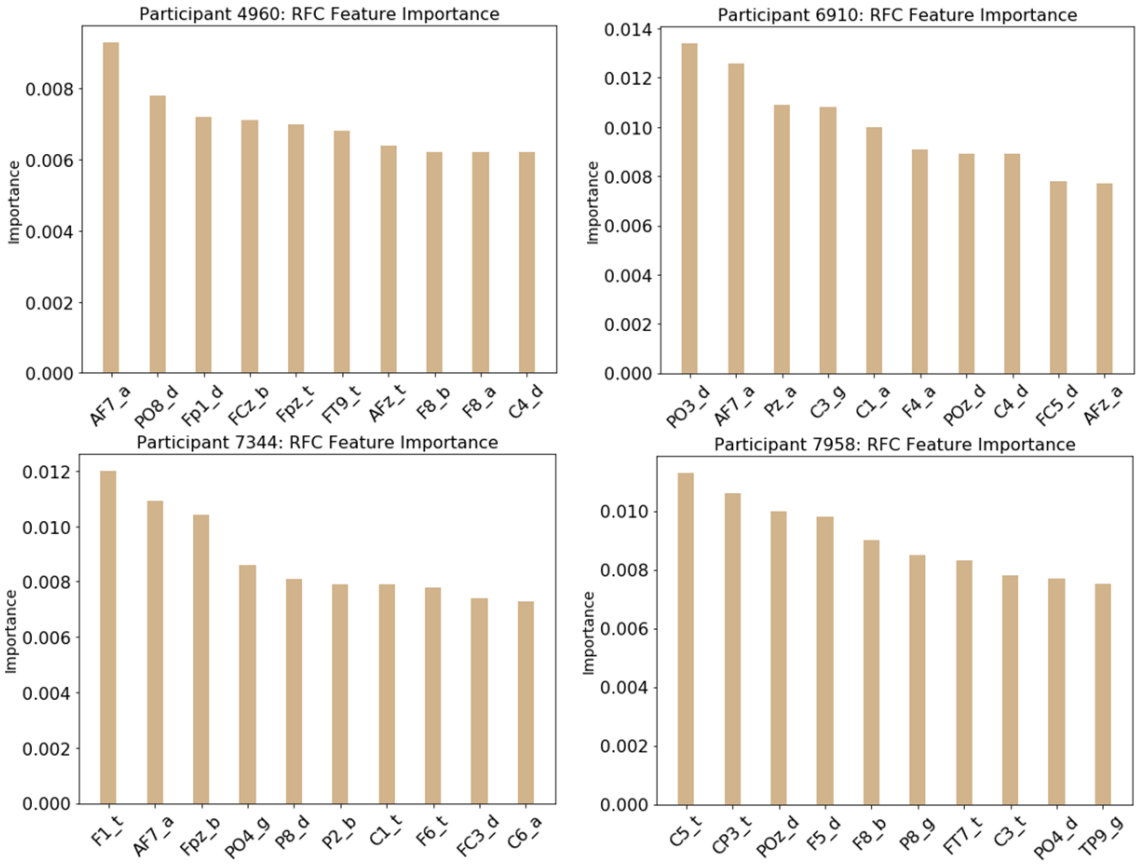

Figure 57: Top Ten RFC Salient Features

4.3.2.2.2 Frequency Analysis

Event-related spectral perturbations (ERSP's) show the change in amplitude of the EEG frequency spectrum as function of time [53]. Minas et al. associated a decrease in log power between 8 -15 Hz with the presence of confirming information [10]. Although ERSP's do not directly translate to the machine learning features as it is an average compared to single trials, observing the ERSP's can help provide insight into whether or not the expected brain activity is present on the averaged signal.

To analyze the machine learning performance with frequency features, cross-participant non-phase locked event-related spectral perturbation (ERSP) showing the

difference of confirming information and disconfirming information at the salient feature electrode locations were created (Figure 58). The color scales represent the log power (dB) difference between confirming and disconfirming signal frequencies. Dark blue indicates decreased power in the confirming signal relative to the disconfirming signal whereas dark red indicates increased power in the confirming signal. The frequency scale ranges from 1 to 30 Hz to be consistent with the frequencies utilized as features in the machine learning and the time ranges from -200ms to 400ms as the decrease in alpha log power for confirming information was observed at stimulus onset to 500ms. For the confirming information at the F2 location, there appears to be a decrease in power around 10 Hz from 100ms to 400ms but in the F4 location, there does not appear to be any drastic decrease in confirming power in the 8-15 Hz range. Both the F6 and F8 locations appear to have a decrease in power in the 5 – 10 Hz range from 0ms to 200ms for confirming information. Although there appears to be some trends in the ERSP's that are in line with previous findings, there was no statistical difference between confirm and disconfirm ERSP's (nonparametric statistical test using Monte Carlo permutations with cluster corrections, alpha=0.01). This lack of difference at the cross-participant level is likely due to noisy brain activity. Due to the nature of the ABC test, participants were free to select information when desired which may have resulted in participants deliberation and selection of information to not lineup. If this occurred, the brain activity associated with each information type would not be consistent, resulting in noisy EEG signals.

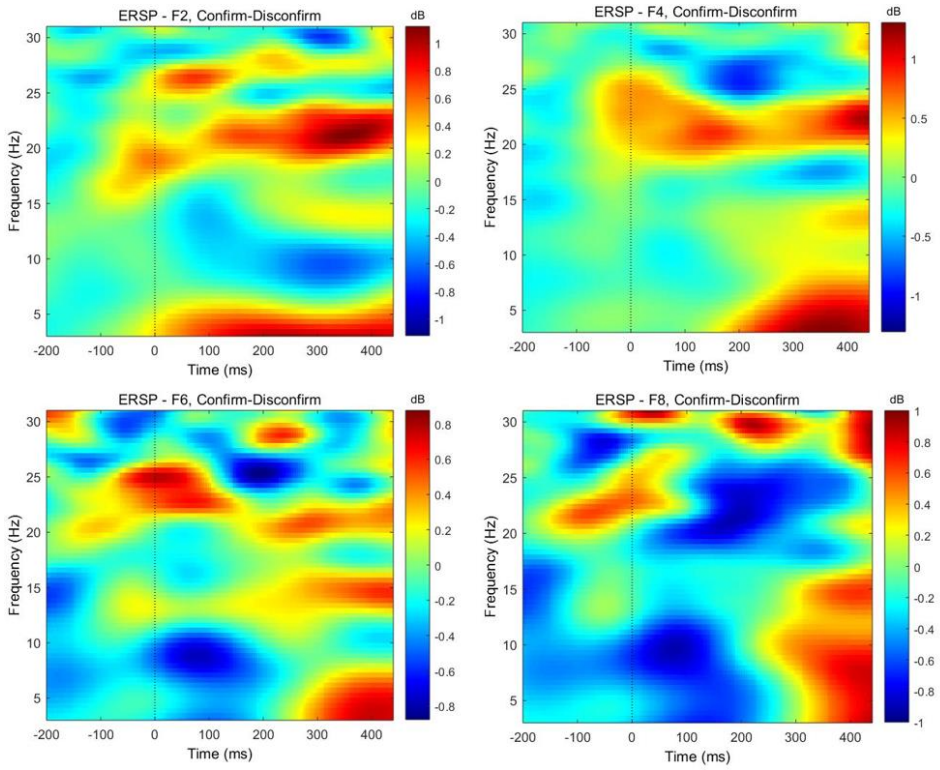

Figure 58: Salient Features Non-Phase Locked ERSP

4.3.2.2.3 Cross-Participant

The results reported in the previous sections of this chapter are within-participant cross-validation results as described in Section 3.4. Since the Frequency Features per Information Selection dataset had the best model performance, cross-participant models were explored on this dataset as outlined in Section 3.4.3.2.1. This dataset had 994 observations across all participants in which 62% is the positive class or the "confirm" label. A train, validation, test approach as applied in which models were trained on 12 participants, validated with 2 participant's data and a single participant was used as a test dataset. This process was repeated so that each participant was the test set. The ensuing results are performance metrics on the test dataset.

The balanced accuracy for cross-participant test performance is displayed in Figure 59. The participant number along the horizontal axis is the participant which was utilized as the test set. A balanced accuracy above the baseline 50% was achieved on 12 of the 15 participants by at least one model. The mean test balanced accuracy (with a 95% confidence interval) for the LDA, RFC and ANN models was 0.496 (± 0.013), 0.493 (± 0.022) and 0.507 (± 0.021). Although above 50% balanced accuracy was obtained on 12 participants, the mean test balanced accuracy was not significantly greater than 50% balanced accuracy (95% confidence interval). The highest balanced accuracy was 0.589 and was obtained by the ANN on participant 6910.

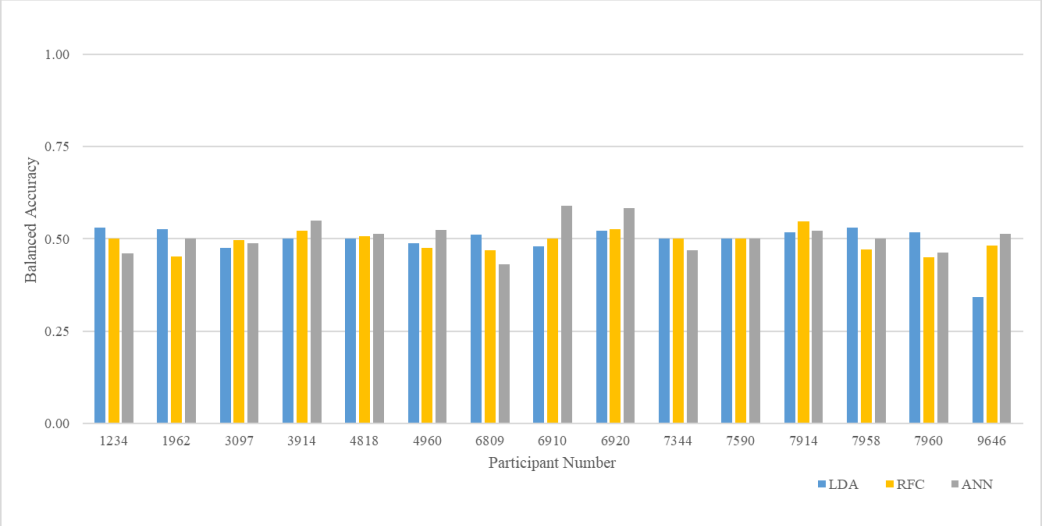

Figure 59: Cross-Participant Model Balanced Accuracy

To compare the cross-participant model performance with the within-participant model performance, the AUROC was examined. The mean cross-participant AUROC for LDA, RFC, and the ANN were 0.489, 0.493, and 0.498 respectively. The cross-participant model AUROC was marginally smaller than the within-participant AUROC

for the LDA, RFC and ANN which was 0.493, 0.485, and 0.521 respectively. The cross-participant AUROC for each participant as the test dataset is displayed in Figure 60.

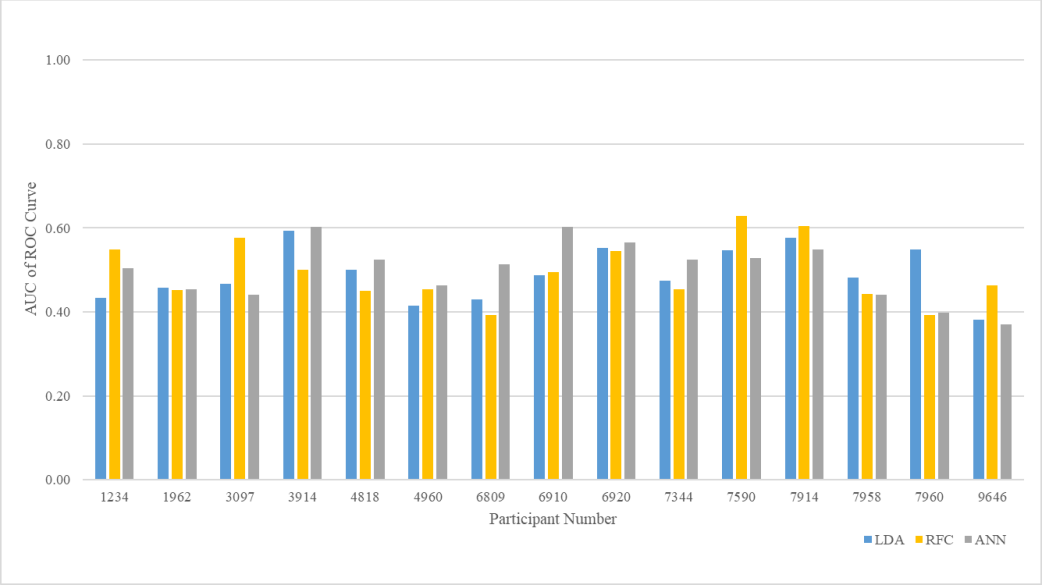

Figure 60: Cross-Participant Model AUROC

The highest AUROC obtained on a single test participant was 0.628 and was obtained by the RFC on participant 7590. These results indicate that models can perform well on the dataset, but the lack of consistent results across all test participants may be due to a multitude of factors. In addition to the small amount of data per participant, one factor which could be affecting the poor cross-participant performance is some participants may have noisy data. For example, as shown in Figure 54, all within-participant models performed poorly on participant 1962 with an AUROC below 0.31. When data is combined for cross-participant models, this noisy data can reduce performance.

4.4 Error Analysis

One hypothesized source of error in the model classification for the Frequency Features per Information Selection dataset was the task type from which the confirming

and disconfirming information was selected from. This dataset consists of information from all of the tasks and if models could not generalize across the different types of tasks the combination of tasks could be a source of the low model performance. To analyze the classification errors contributed to each task, classification errors across all participants were summed. The number of observations from each task which contributed to the total observation classification error was then determined as a percentage of the total classification error. The task error distribution for the RFC is displayed in Table 14.

Table 14: Task Classification Error Distribution for RFC

Task	Percent of all observations (%)	Percent of observations with errors (%)
Stand	19.8	25.5
Comparison	27.3	31.2
Intel	18.3	14.8
HR	30.7	32.5

While only the RFC error analysis is displayed, all classifiers had nearly identical results. The RFC performed the best on the Intel tasks and the worst on the Stand tasks with the Intel task having a lower distribution of observations in the errors and the Stand task having a higher distribution. Although the errors vary by task, the task distribution of classification errors is fairly close to the task distribution among all observations. These results suggest the task source of the information is not a main cause of the classification errors.

4.5 Summary

Contrary to the hypothesized effects that decision tasks with an initial decision would result in more unbalanced information search, they actually appeared to result in more balanced information search. Stand tasks with an initial decision resulted in a more

unbalanced information search compared to Stand tasks without an initial decision. This trend could be due to participants experiencing cognitive dissonance after making an initial decision.

The behavioral patterns of information selection or information/question importance were used to quantify the presence of confirmation bias in a task. These behaviors were used to measure a participant's level of bias on a task by finding the proportion of selected confirming information. Both task completion time and information search completion time behavior patterns were explored for association with biased behavior. In the early data exploration stages, a significant trend was observed in the task completion times of the Stand, Comparison and HR tasks. The trend resulted in reduced completion time as the order of the task increased in the ABC assessment. In addition, the information search completion time of the Stand and Comparison tasks were also affected by this decreasing trend in completion time. For these reasons, further association with completion time and confirmation were not explored. A participant's information revisits was also explored for association with biased behavior but no significant trends were observed.

The machine learning balanced accuracy metrics for all four explored datasets are displayed in Table 15. The first four columns of results are within-participant cross-validation results, while the final column is cross-participant test results. Each column corresponds to the results for the highest performing model for the respective dataset. The two datasets associated with segmenting the data by task performed rather poorly. The Frequency Features per Information Selection had the highest model performance, achieving model performance significantly greater than 50% baseline on participants

6809 and 7960. The cross-participant models obtained test performance greater than baseline on four participants, but the mean test accuracy across all participants was not significantly greater than baseline.

Table 15: Machine Learning Summary of Balanced Accuracy Results

Participant ID	Time Series per Task	Frequency per Task	Time Series per Info.	Frequency per Info.	Cross-Participant Frequency per Info.
1234	-	-	↑	-	-
1962	-	-	↑	-	-
3097	-	-	-	-	-
3914	-	-	↑	↑	↑
4818	-	-	-	-	↑
4960	-	↑	↑	↑	-
6809	-	-	-	↑*	-
6910	↑	-	-	↑	-
6920	↑	↑	↑*	↑	↑
7344	-	-	-	-	-
7590	-	-	-	-	-
7914	-	-	↑	-	↑
7958	-	-	-	↑	-
7960	-	↑	-	↑*	-
9646	-	-	-	-	-

Legend	
-	below 50%
↑	above 50%
↑*	above 50% and significant (if applicable)

The low machine learning performance in the information selection datasets may be a result of the fact that there is no statistical difference between the cross-participant time series signals of confirming and disconfirming information. Despite the lack of significance between the two conditions, the expected brain activity was present in

participant 6809 which had one of the best model performances. In addition, at least one of the expected features were one of the top eight most salient features in four of the participants with the highest RFC performance.

V. Conclusions and Recommendations

5.1 Conclusions of Research

This research was successful in investigating decision-based cognitive bias relationships between behavior, self-reported information and psychophysiological signals collected when a participant conducts a decision affected by confirmation bias. These relationships were investigated by examining significant behaviors associated with a biased task, as well as exploring machine learning methods to detect confirming and disconfirming information in a task. Detecting the processing of confirming and disconfirming information can allow subjective detection of confirmation bias by determining when confirming information is inappropriately sought in decision making. In addition, time series signals and machine learning models' salient features were utilized to explore brain activity associated with confirming and disconfirming information.

The first relationship investigated in this work was the effect of making an initial decision on subsequent information search, specifically in the Stand task. This relationship was investigated to answer research question one (see Section 3.2) and was hypothesized that the absence of an initial decision prior to information search would result in more balanced information search compared to making an initial decision. Contrary to this hypothesis, findings indicate when participants did not make an initial decision prior to information search, the participants information search was less balanced. Across all participants, the Stand tasks without an initial decision appeared to result in a higher proportion of selected confirming information compared to Stand tasks

with an initial decision. In the Intel task the relationship between the presence of an accepted hypothesis and evidence importance was investigated. Across all participants there did not appear to be an effect on evidence importance between Intel tasks with and without an accepted hypothesis.

Research question two was focused on quantifying a participant's level of confirmation bias in each ABC assessment task and subsequently associating behavior patterns with high levels of bias. The level of bias in the Stand and Comparison tasks were quantified by the proportion of confirming information selected for which the cross-participant mean level of bias was 0.52 and 0.65 respectively. The level of bias in the Intel and HR tasks were quantified by the proportion of confirming evidence and questions selected for which the cross-participant mean level of bias was 0.71 and 0.59 respectively. Behavior patterns in task and information search completion time were investigated for association with levels of bias in each task. Unfortunately, task order had a significant impact on completion time and thus was not investigated further for association with confirmation bias. Lastly, patterns in information revisits were also explored for association with bias. But there were no significant patterns across all participants. Regarding this focus of research, confirmation bias was successfully quantified for each task but no significant patterns in behaviors were associated with the level of bias.

To investigate research question three (see Section 3.2) multiple machine learning avenues were explored to determine if a machine learning classification model can detect the presence of confirming and disconfirming information with performance greater than random chance. Machine learning models were able to obtain a within-participant cross-

validation balanced accuracy above 50% on 10 of the 15 participants, but only two participants were significantly greater. Despite the overall low performance, the highest balanced accuracy on a single participant was only 62.6% which was obtained on participant 6809 by the ANN. In addition, the highest within-participant area under the ROC for a model across all participants was 51.2% and was obtained by the ANN. Cross-participant machine learning was also explored on the Frequency Features per Information Selection Dataset to increase the amount of training data and allow model tuning. Model test balanced accuracy was greater than 50% on four participants, with the highest achieved test balanced accuracy being 58.9% obtained by the ANN on participant 6910. But none of the three explored model's mean cross-participant balanced accuracy was significantly greater than 50%, indicating performance was not consistent across participants. Overall machine learning model performance was rather low. Further machine learning exploration with a larger within-participant dataset is necessary to determine the highest achievable test performance with model tuning.

Two methods were employed in this work to determine if neurophysiological signals in the right frontal lobe are associated with confirming and disconfirming information: machine learning salient features, and averaged time series signals. Salient features in the random forest classier were explored, but no features associated with brain activity in the right frontal lobe appeared to be significant in all 15 participants. Despite the lack of feature consistency across all participants, at least one feature associated with the expected brain location was among the eight most important features in four participants. These four participants were also among the participants with the highest random forest model performance. These results indicate a random forest model which

uses a feature associated with the right frontal lobe of the brain is likely to have higher model performance than a model which does not use such features. Comparison of cross-participant time series signals for confirming and disconfirming information also had similar results. There was no significant difference in the cross-participant time series signals of the two types of information at EEG locations associated with the brains right frontal region. But within-participant time series signals of the participant with the highest model performance appeared to have increased brain activity in the presence of confirming activity. Although no conclusions can be drawn at the cross-participant level, participants with high model performance appear to have some of the expected brain activity in time series signals and salient features associated with the expected brain activity.

5.2 Significance of Research

Traditional methods of estimating confirmation bias employ subjective, self-report methods to measure a decision-task with bias. These methods are not suitable for estimating bias in real-time due to the inherent delay and subjectivity of the measures. With the crucial effect biases can have in military operations, a more robust, dynamic method of estimating bias is necessary. The results in this work support the potential for objective measurement of confirmation bias through neurophysiological measures. The machine learning models achieved within-participant performance above the baseline balanced accuracy of 50% on 2 of the 15 participants, which indicates classifying confirming and disconfirming is feasible, but did not meet this works objective of at least three participants above 50%. In addition, model features associated with brain activity in

right frontal lobe were one of the eight most salient features in four participants. These results indicate the machine learning models could be using the expected brain activity to differentiate between confirming and disconfirming information. However, not all participants had consistent machine learning results and prominent features associated with the expected brain region. Future work is necessary to generate more data observations for machine learning with a closer focus on classifying information as confirming or disconfirming using frequency features.

5.3 Recommendations for Future Research

5.3.1 ABC Assessment Changes

For this work, minimal modifications were made to the ABC assessment to prevent tampering with literature-backed decision tasks (see Section 3.3 for implemented assessment modifications). In the early stages of data exploration, it was apparent that the assessment was not optimal for physiological data collection or machine learning. The data generated from the ABC assessment was not ideal for machine learning because of the small number of observations generated. The maximum information that could have been selected by a participant is 144, with the actual number ranging from 42 to 93 across the participants. For a within-participant model the number of observations per participant is insufficient for a proper train, validation, test data split.

One possible approach to generate more observations is to either double the number of decision tasks in the ABC assessment. This would likely double the number of observations to approximately 84 to 186 observations per participant. The disadvantage of doubling the number of decision tasks is the subject participation time would double

and fatigue could produce poor physiological results. A different method to improve the number of observations which should also be explored is to only use one type of decision task for all 14 decision tasks. This would allow a biased and unbiased decision task to be used in a train and test set for machine learning.

Another approach to generate more observations is to have participants complete a pre-assessment that could provide more observations. A possible approach is to generate 150 sentences on controversial topics which confirm or disconfirm a stance on the given topic. The pre-assessment would present one sentence at a time for ten seconds to the participant. After the ten seconds, the participant would have five seconds to determine if they agree, disagree, or are neutral regarding the sentence. The pre-assessment approach would generate 150 observations (labeled as confirming, disconfirming or neutral) while only adding approximately 37.5 minutes to participation time. One potential disadvantage with the pre-assessment, is the brain activity present could be different than the brain activity in a complex decision task. But as long as the brain activity is similar enough with the same activation from confirming information it would provide more observations with minimal added participation time.

The structure of the decision tasks in the ABC assessment was not optimal for physiological data collection because it allowed participants to read information without selecting it and participants could quickly click through information. The research investigator observed participants during the assessment and some general trends were noticed. First, on the Comparison tasks some participants displayed comments by clicking through all comments at once, despite the goal of minimizing the number selected comments. When this occurred, the participant's deliberation of each piece of

information was not in sync with when they selected the comment. These observations were removed from the machine learning datasets making an already small dataset even smaller. Secondly, in the Intel and HR tasks participants generally read each question/evidence and then made their selections all at once. Since the electrophysiological data was time-locked to when the participant selected the questions/evidence the brain activity during the deliberation may have been missed.

To prevent the misalignment of the participant's information deliberation and the time-stamped physiological data, only one piece of information in a task should be displayed at a time. The ABC assessment would display how many pieces of information are available for a given task, but display only one piece of information until the participant chooses to proceed to the next piece of information. Once every piece of information is reviewed by the participant, the participant could freely review all information. Structuring the decision tasks in this manner would ensure the participant is only deliberating on one piece of information during the time that piece of information is displayed. A shortcoming to this structure is the decision task would not mimic information search in a real-world decision. Despite this limitation, the changed structure may show greater machine learning performance is achievable prior to replication in a near real-world decision task.

Finally, because the earliest task took far longer than subsequent tasks, and because the first task of each type took longer than the remaining tasks of each type, a training session for each of the task types prior to the experiment would help reduce the effect of not being familiar with the task on the true duration of the participant's activity in assessing information and making decisions. Additionally, to de-trend the way

learning effect perturbs task completion time, the task order within the ABC assessment should be counterbalanced. Changing the order of presented tasks for each participant would determine if inter-task differences in completion time are due unfamiliarity or because of the differences in duration of the participant's behavior task within the task. Implementing this change could allow completion time to be associated with confirmation bias.

5.3.2 Machine Learning

With the limited number of observations per participant, the machine learning methods in this work were cursory and were mainly used to determine if the problem was suitable for machine learning. With a proper sized dataset, a train, validation and training set should be used for model tuning. Utilizing model tuning, multiple machine learning models could be compared to show if one model is best for the problem domain. In addition, with more complex, tuned models, a higher model performance is likely achievable. Thus, future work should explore generating a new dataset with a significant increase in observations.

While a larger dataset would likely improve performance, there are still multiple machine learning facets that can be explored further on these datasets. This work focused on mean power spectral density of the clinical frequencies and the raw time series signals as features. Recent research on cognitive workload estimation from EEG has shown variance of power spectral density as a significant feature [54]. Consequently, using variance in addition to the mean power spectral density may show improvement in performance. Another avenue that could be explored is the approach utilized on the task

dataset. This work implemented a classification approach to estimate a decision-task as biased. An alternative approach could use a regression problem in which the proportion of confirming information is estimated. The regression approach could improve performance by predicting the actual proportion of confirming information instead of the simplified biased or unbiased method in this work.

5.3.3 Physiological Measures

Although this work focused on EEG analysis, EOG and ECG were also collected. Analysis of these physiological measures could show relationships with confirmation bias. In addition, galvanic skin response (GSR) should be collected in future work. Increased arousal measured through GSR at six seconds after information onset has been associated with hypothesis confirming information compared to disconfirming information [10]. Unfortunately, this work had no GSR-sensing equipment. Associating increased arousal with presence of decision-confirming information, relative to disconfirming information, may help quantify the presence of information confirming the participant's belief. In addition, operationalizing the participant's emotional response when they select confirming and disconfirming in thel information search portion of the ABC assessment could be an added feature in the machine learning models. A greater overall emotional response to information correlated with increased brain activity of the right frontal portion of the brain could boost machine learning performance immensely. However, some decision task design changes to the current ABC assessment may be necessary to capture proper GSR responses. Currently, the ABC assessment allows participants to freely select information at any time. Although, this method closely

mirrors a complex decision, it is not suitable for ensuring a six second GSR response can be associated with selected information. Participants could select multiple pieces of information within a six second window which could cause GSR responses to overlap. For recommended task changes to improve physiological data collection, see Section 5.3.1.

In addition to the aforementioned measures, utilizing functional near-infrared spectroscopy (fNIRS) to associate brain activity with confirmation bias measures should also be explored. fNIRS is a non-invasive optical imaging technique which measures blood flow response to brain activity [55]. fNIRS tends to provide better spatial but lower temporal resolution than EEG and can also be less susceptible to noise artifacts [56]. fNIRS has been utilized to associate activation in the dorsolateral prefrontal cortex with working memory [57] and has shown promise for investigating decision making [58]. Utilizing fNIRS in future work could augment collected EEG with better spatial information that could be beneficial in classifying information.

5.3.4 Participant Selection for Future Trials

As noted in the limitations (Section 1.6.2) the participant demographics for this work were not diverse. To validate the results of this work on a larger scale, a more diverse participant pool should be solicited. Both males and females and a wider range of ages, backgrounds, education levels, and diversity in other factors should be included. Given the goal of this work is estimating confirmation bias in general decision-tasks a specific population is not necessary, but this work could also be replicated with a targeted population. This would determine if there are specific behaviors or physiological

measures associated with confirmation bias in a specific task like intelligence analysis or cyber defense analysis.

5.4 Summary

This work explored estimating confirmation bias in decision making through subjective measures by classifying EEG signals from decision-confirming or disconfirming information. The machine learning performance objective measured by balanced accuracy was not met as only 2 of the 15 participants obtained above 50% balanced accuracy, indicating subjective estimation of confirmation bias may be feasible but more work is necessary. In addition, features associated with brain locations that have been related to the presence of confirming information, were salient features for participants with the highest model performance. But the lack of consistent performance across participants indicate experiment design changes are necessary for improved performance. Experiment design changes are not only necessary for improved physiological signal collection, but also so information acquisition behavior patterns can be associated with confirmation bias. If the explored confirmation bias estimation performance can be improved, suboptimal decisions due to confirmation bias could be detected and prevented.

Appendix A: Balanced Accuracy Incorrect Significance

The binomial probability formula below gives the probability of exactly *k* success in *n* trials [59]:

$$C(n,k)\ p^k\ q^{n-k}$$

n: number of independent trials

k: number of successes

p: probability of success

q: probability of failure $(1 - \mathrm{p})$

With an alpha of 0.05 in a 95% confidence interval, there is a 5% probability of obtaining incorrect significance by random chance on any given participant. Using the binomial probability formula above, with 15 participants there is a 46.3% chance of obtaining incorrect significance on exactly zero participants, 36.6% chance on exactly one participant, and a 13.5% chance on two participants. The probability of obtaining incorrect significance on more than two participants is one minus the sum of these probabilities and is 3.6%. Therefore more than two participants must be significantly above 50% balanced accuracy to obtain a significance with an alpha of 0.05.

Appendix B: IRB Approval Letter

DEPARTMENT OF THE AIR FORCE
AIR FORCE RESEARCH LABORATORY
WRIGHT-PATTERSON AIR FORCE BASE OHIO 45433

MEMORANDUM FOR AFIT (BRETT BORGHETTI)

FROM: 711 HPW/IR

SUBJECT: IRB Approval for the Use of Human Volunteers in Research

1. Protocol title: Cognitive Bias Estimation in Cyber Security

2. Protocol number: FWR20180174H

3. Protocol version: v1.00

4. Risk: Minimal

5. Approval date: 07 September 2018

6. Expiration date: 06 September 2019

   Your renewal submission date is ***one month prior*** to your expiration date. The renewal is due 06 August 2019

7. Review Category: 32CFR219.110 (b)(4) & (b)(7)

8. We have confirmed that all individuals involved in this study are covered under a valid Assurance.

9. The study objective is identify statistical relationships, and computationally model the associations between three components: Self-reported measures of confirmation bias in decision-making; Behavior patterns during investigative decision-making; Physiological signals collected during decision-making.

10. A waiver of documentation of consent has been granted for this research project as it meets the criteria outlined in 32 CFR 219.117 (c).

11. All inquiries and correspondence concerning this protocol should include the protocol number and name of the primary investigator. Please contact the 711 HPW/IR office using the organizational mailbox at AFRL.IR.ProtocolManagement@us.af.mil or calling 937-904-8094 [DSN 674].

LONDON.KIM.ELIZABETH.1155556370

Digitally signed by LONDON.KIM.ELIZABETH.1155556370
Date: 2018.09.07 11:57:46 -04'00'

KIM E. LONDON, JD, MPH
Chair, AFRL IRB

Appendix C: Abbreviated Informed Consent Document

Abbreviated Informed Consent Document
Cognitive Bias Estimation in Cyber Security
FWR20180174

You are being asked to participate in a research study. The purpose of the study is to study the relationships between decision confidence, brain signals and patterns of investigative behavior.

The expected length of your participation is up to 2 hours on two separate days within a two week period (total of 4 hours).

If you participate in this research, you will be performing decision-making tasks on a computer. Additionally, an Electroencephalograph (EEG) head cap will be applied to measure brain activity. Sensors placed near your eyes for Electrooculography (EOG) will measure eye movement and blink signals while sensors on your chest will record heart information using Electrocardiography (ECG). Galvanic Skin Response (GSR) will measure electrodermal activity (EDA) and will be placed on fingers of your non-mouse hand.

The computerized portion of this study does not involve any more than minimal risk to you. In other words, there is no harm or discomfort beyond what is ordinarily encountered in daily life when using the computer or during the performance of routine physical or psychological tests.

There are other possible sources of risk & discomfort. We will attach sensors on your head, face and arms. Some participants may experience discomfort (due to limited movement during trials). Minor skin irritation and/or discomfort may result when the electrodes are placed the head, face, chest and abdomen when you or the testers clean those locations to reduce electrical impedance in order to improve signal quality. Minimal, temporary hair loss is unlikely but may occur locally at the electrode sites. Because applying the electrical sensors to participants requires making contact with and, in some cases, scrubbing the skin with liquids and exfoliating materials, there is a theoretical risk of transmitting skin-borne pathogens during this process. All necessary equipment will be cleaned and disinfected before the procedure begins to minimize the risk of transmitting diseases. All equipment is standard and is used commercially.

If you choose to fill out the questionnaire, steps will be taken to protect your confidentiality as described below.

You are not expected to benefit directly from this research.

Your decision to participate in this research is voluntary. You can discontinue participation at any time without penalty or loss.

The researchers will take the following precautions to maintain the confidentiality of your data: The data collected from your questionnaire, computer activities and the sensors on your body will be associated with a randomly-assigned participant number, but no personal information will be collected with this data. The un-identifiable data will be protected with a password on the collection computers and via CAC access on the AFIT network. The researchers will/will not collect any identifiers linked to you. No participant identifiable information will be included in any publications. Any paper data collected will be kept in a locked cabinet.

The data may be accessed by the Department of Defense for auditing purposes.

If you have questions regarding the study, contact the Principal Investigator: Dr. Brett Borghetti can be reached at (937) 255-3636x4612. If you have questions regarding your rights as a research subject, contact the AFRL IRB: 937-904-8100 or afrl.ir.protocolmanagment@us.af.mil.

Appendix D: Pre-Experiment Questionnaire

ID: _____ Date: _____

Pre-Experiment Questionnaire (ONLY Experiment Day)

How many hours of sleep did you have last night?
Circle one choice: 0-4 hours, 5-6 hours, 7-9 hours, 9+ hours

How would you characterize your sleep last night?
Circle one choice: Very Poor, Poor, Fair, Good, Very Good

Did you consume any products with caffeine today?
Circle one choice: yes or no
If yes:
What product(s) did you consume?

_____
When did last consume this product?

_____
Approximately how much (mg / ounces / cups) of this product have you consumed today? _____

Do you have any reason(s) to believe that your ability to accomplish tasks during this study today would be abnormal (for example: distracted, overly tired, hungry, stressed, injured)?

_____

_____

If yes:
Do you still want to participate in the cyber study today? Circle one choice: Yes / No
If no:
Would you like to reschedule participation for another day? _____

Appendix E: Post-Experiment Questionnaire

ID: _____ Date: _____

Post-Experiment Questionnaire (ONLY Experiment Day)

Computer experience:

What sort of electronic devices do you use?

Circle all choices:

Personal computer/Desktop/Laptop
TV/Game Console
Smartphone/Tablet
Enterprise Server
Other, _____

How often do you use electronic devices?

Circle one choice: Daily, A few times a week, Once a week, Never

Do you use electronic devices in your job?

Circle one choice: Yes, No, Prefer not to answer

Age: _____

Are you male or female? Male___ Female____ Prefer not to answer ____

What's your highest education level?

A. Lower than high school
B. Graduated from high school
C. Some college, no degree
D. Associate's Degree
E. Bachelor's Degree
F. Master's degree
G. Ph.D. degree

Bibliography

[1] A. Tversky and D. Kahneman, “Judgment under uncertainty: Heuristics and biases,” *Science*, vol. 185, pp. 1124–1131, 1974.

[2] National Research Council, “Measuring Human Capabilities: An Agenda for Basic Research on the Assessment of Individual and Group Performance Potential for Military Accession. :,” The National Academies Press, Washington, DC, 2015.

[3] J. R. Gersh and N. Bos, “Cognitive and Organizational Challenges of Big Data in Cyber Defense,” *Proc. 2014 Work. Hum. Centered Big Data Res. - HCBDR ’14*, pp. 4–8, 2014.

[4] R. S. Nickerson, “Confirmation bias: A ubiquitous phenomenon in many guises.,” *Rev. Gen. Psychol.*, vol. 2, no. 2, pp. 175–220, 1998.

[5] H. H. Huang, J. S. C. Hsu, and C. Y. Ku, “Understanding the role of computer-mediated counter-argument in countering confirmation bias,” *Decis. Support Syst.*, vol. 53, no. 3, pp. 438–447, 2012.

[6] M. A. Tolcott, F. F. Marvin, and P. E. Lehner, “Expert Decisionmaking in Evolving Situations,” *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 3, pp. 606–615, 1989.

[7] M. E. Oswald and S. Grosjean, “Confirmation bias,” *Cogn. Illusion. A Handb. Fallacies Biases Thinking, Judgement Mem.*, no. August, pp. 79–96, 2004.

[8] R. E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences*, Third. Brooks/Cole Publishing Company, 1995.

[9] A. Nussbaumer, K. Verbert, E. C. Hillemann, M. A. Bedek, and D. Albert, “A framework for cognitive bias detection and feedback in a visual analytics

environment," *Proc. - 2016 Eur. Intell. Secur. Informatics Conf. EISIC 2016*, pp. 148–151, 2017.

[10] R. K. Minas, R. F. Potter, A. R. Dennis, V. Bartelt, and S. Bae, "Putting on the Thinking Cap: Using NeuroIS to Understand Information Processing Biases in Virtual Teams," *J. Manag. Inf. Syst.*, vol. 30, no. 4, pp. 49–82, 2014.

[11] M. J. Janser and N. D. Wilson, "Cognitive Biases in Military Decision Making," *USAWC Cl. 2007*, pp. 1–16, 2007.

[12] D. C. Gompert and R. L. Kugler, "Custer in Cyberspace," *Def. Horizons*, vol. 51, no. February, pp. 1–11, 2006.

[13] J. E. (Hans) Korteling, A.-M. Brouwer, and A. Toet, "A neuroscientific perspective on cognitive biases," *Open Sci. Framew.*, 2017.

[14] E. Shafir and R. LeBoeuf, "Rationality," *Annu. Rev. Psychol.*, vol. 53, pp. 491–517, 2002.

[15] M. Zeleny, "Multiple Criteria Decision Making (MCDM): From Paradigm Lost to Paradigm Regained?," *J. MultiCriteria Decis. Anal.*, vol. 18, pp. 77–89, 2011.

[16] M. Fleischmann, M. Amirpur, A. Benlian, and T. Hess, "Cognitive Biases in Information Systems Research: a Scientometric Analysis," *ECIS 2014 Proc.*, no. 26, pp. 1–21, 2014.

[17] F. Bacon, *Novum Organum*. New York: Collier P.F., 1902.

[18] P. C. Wason, "On the failure to eliminate hypotheses in a conceptual task," *Q. J. Exp. Psychol.*, vol. 12, no. 3, pp. 129–140, 1960.

[19] J. Klayman and Y. Ha, "Confirmation, disconfirmation, and information in hypothesis testing.," *Psychol. Rev.*, vol. 94, no. 2, pp. 211–228, 1987.

[20] M. Snyder and W. Swann, “Hypothesis-testing processes in social interaction.,” *J. Pers. Soc. Psychol.*, vol. 36, no. 11, pp. 1202–1212, 1978.

[21] K. Ask and P. A. Granhag, “Motivational sources of confirmation bias in criminal investigations: the need for cognitive closure,” *J. Investig. Psychol. Offender Profiling*, vol. 2, no. 1, pp. 43–63, 2005.

[22] C. G. Lord, L. Ross, and M. R. Lepper, “Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence,” *J. Pers. Soc. Psychol.*, vol. 37, no. 11, pp. 2098–2109, 1979.

[23] M. B. Cook and H. S. Smallman, “Human Factors of the Confirmation Bias in Intelligence Analysis: Decision Support From Graphical Evidence Landscapes,” *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 50, no. 5, pp. 745–754, 2008.

[24] J. Evans, *Bias in Human Reasoning: Causes and Consequences*. Lawrence Erlbaum Associates, 1989.

[25] A. Tversky, “Features of similarity,” vol. 84, no. 4, 1977.

[26] C. R. Mynatt, M. E. Doherty, and R. D. Tweney, “Confirmation bias in a simulated research environment: An experimental study of scientific inference,” *Q. J. Exp. Psychol.*, vol. 29, no. 1, pp. 85–95, 1977.

[27] E. Jonas, S. Schulz-hardt, D. Frey, and N. Thelen, “Confirmation Bias in Sequential Information Search After Preliminary,” *J. Pers. Soc. Psychol.*, vol. 80, no. 4, pp. 557–571, 2001.

[28] P. E. Lehner, L. Adelman, B. A. Cheikes, and M. J. Brown, “Confirmation bias in complex analyses,” *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*, vol. 38, no. 3, pp. 584–592, 2008.

[29] M. X. Cohen, *Analyzing Neural Time Series Data: Theory and Practice*. 2014.

[30] J. Braithwaite, D. Watson, J. Robert, and R. Mickey, "A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments," …, pp. 1–42, 2013.

[31] M. E. Dawson, A. M. Schell, and C. G. Courtney, "The Skin Conductance Response, Anticipation, and Decision-Making," *J. Neurosci. Psychol. Econ.*, vol. 4, no. 2, pp. 111–116, 2011.

[32] A. Geron, *Hands-On Machine Learning with Scikit-Learn & TensorFlow*, 1st ed. Sebastopol: O'Reilly, 2017.

[33] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, 7th Editio. New York: Springer, 2017.

[34] B. Binias, D. Myszor, and K. A. Cyran, "A Machine Learning Approach to the Detection of Pilot ' s Reaction to Unexpected Events Based on EEG Signals," *Comput. Intell. Neurosci.*, vol. 2018, 2018.

[35] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks," pp. 1–15, 2015.

[36] D. R. Edla, K. Mangalorekar, G. Dhavalikar, and S. Dodia, "Classification of EEG data for human mental state analysis using Random Forest Classifier," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1523–1532, 2018.

[37] F. Chollet, *Deep Learning with Python*, no. 1. 2017.

[38] G. Gao, L. Shang, K. Xiong, J. Fang, C. Zhang, and X. Gu, "EEG Classification Based on Sparse Representation and Deep Learning," *2017 Int. Jt. Conf. Neural Networks*, vol. 16, no. 6, pp. 789–795, 2018.

[39] S. Bai, J. Z. Kolter, and V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” 2018.

[40] E. Popov and S. Fomenkov, “Classification of Hand Motions in EEG Signals using Recurrent Neural Networks,” *2016 2nd Int. Conf. Ind. Eng. Appl. Manuf.*, pp. 0–3, 2016.

[41] R. Hefron, B. Borghetti, C. S. Kabban, J. Christensen, and J. Estepp, “Cross-participant EEG-based Assessment of Cognitive Workload Using Multi-Path Convolutional Recurrent Neural Networks,” *Sensors (Switzerland)*, vol. 18, no. 5, 2018.

[42] A. Gertner, F. Zaromb, R. Schneider, R. D. Roberts, and G. Matthews, “The Assessment of Biases in Cognition,” 2016.

[43] P. Fischer, E. Jonas, D. Frey, and A. Kastenmüller, “Selective exposure and decision framing: The impact of gain and loss framing on confirmatory information search after decisions,” *J. Exp. Soc. Psychol.*, vol. 44, no. 2, pp. 312–320, 2008.

[44] MITRE, “IARPA Sirius Program Assessment of Biases in Cognition (ABC),” 2015.

[45] E. Maris and R. Oostenveld, “Nonparametric statistical testing of EEG- and MEG-data,” *J. Neurosci. Methods*, vol. 164, pp. 177–190, 2007.

[46] S. Raschka, “Predictive Modeling, supervised machine learning, and pattern classification - the big picture,” 2014. [Online]. Available: https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html. [Accessed: 05-Sep-2018].

[47] A. Delorme and S. Makeig, “EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis,” *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[48] M. Miyakoshi, “Makoto’s Preprocessing Pipeline,” *Swartz Center for Computational Neuroscience*, 2018. [Online]. Available: https://sccn.ucsd.edu/wiki/Makoto’s_preprocessing_pipeline. [Accessed: 21-May-2018].

[49] M. B. Pontifex, V. Miskovic, and S. Laszlo, “Evaluating the efficacy of fully automated approaches for the selection of eyeblink ICA componenets,” *Psychophysiology*, vol. 54, pp. 780–791, 2017.

[50] F. Lotte, M. Congedo, and L. Anatole, “A review of classification algorithms for EEG-based brain – computer interfaces To cite this version : A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces,” 2007.

[51] O. Ledoit and M. Wolf, “Honey, I Shrunk the Sample Covariance Matrix,” *J. Portf. Manag.*, vol. 30, no. 4, pp. 110–119, 2004.

[52] S. Sur and V. Sinha, “Event-related potential: An overview,” *Ind. Psychiatry J.*, vol. 18, no. 1, pp. 70–73, 2009.

[53] S. Makeig, “Auditory Event-Related Dynamics EEG Spectrum and Effects of Exposure to Tones,” *Electroencephalogr. Clin. Neurophysiol.*, no. 89, pp. 283–293, 1993.

[54] R. G. Hefron, B. J. Borghetti, J. C. Christensen, and C. M. S. Kabban, “Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation,” *Pattern Recognit. Lett.*, vol. 94, pp. 96–104, 2017.

[55] V. Scarapicchia, C. Brown, C. Mayo, and J. R. Gawryluk, “Functional Magnetic Resonance Imaging and Functional Near-Infrared Spectroscopy: Insights from Combined Recording Studies,” *Front. Hum. Neurosci.*, vol. 11, no. August, pp. 1–12, 2017.

[56] J. Shin, A. Von Lühmann, D. W. Kim, J. Mehnert, H. J. Hwang, and K. R. Müller, “Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset,” *Sci. Data*, vol. 5, pp. 1–16, 2018.

[57] I. L. Kwee and T. Nakada, “Dorsolateral prefrontal lobe activation declines significantly with age: Functional NIRS study,” *J. Neurol.*, vol. 250, no. 5, pp. 525–529, 2003.

[58] I. M. Kopton and P. Kenning, “Near-infrared spectroscopy (NIRS) as a new tool for neuroeconomic research,” *Front. Hum. Neurosci.*, vol. 8, no. August, pp. 1–13, 2014.

[59] K. H. Rosen, *Discrete Mathematics and Its Applications*. McGraw-Hill, 2011.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
21-03-2019	Master's Thesis	September 2017 – March 2019

TITLE AND SUBTITLE

Confirmation Bias Estimation from Electroencephalography with Machine Learning

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Villarreal, Micah, Captain, USAF

5d. PROJECT NUMBER
19G192

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENG-MS-19-M-065

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
Dr. James Lawton
AFOSR Program Manager, Information and Networks
875 N. Randolph Street, Suite 325, Arlington, VA 22203-1768
(703) 696-5999
james.lawton.1@us.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR/RTA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT
Cognitive biases are known to plague human decision making and can have disastrous effects in the fast-paced environments of military operators. Traditionally, behavioral methods are employed to measure the level of bias in a decision. However, these measures can be hindered by a multitude of subjective factors and cannot be collected in real-time. This work investigates enhancing the current measures of estimating confirmation bias with additional behavior patterns and physiological variables to explore the viability of real-time bias detection. Confirmation bias in decisions is estimated by modeling the relationship between Electroencephalography (EEG) signals and behavioral data using machine learning methods.

15. SUBJECT TERMS
confirmation bias, decision making, Assessment of Bias in Cognition (ABC), behavior patterns, electroencephalography (EEG)

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dr. Brett J. Borghetti, AFIT/ENG
U	**U**	**U**	**UU**	177	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565, ext 4612 brett.borghetti@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18