

Measuring and Modeling of Human Trust in Machine using Brain Signals

Chidi Happy Onyeoru

16th March 2026
Version: Final Version

Measuring and Modeling of Human Trust in Machine using Brain Signals

by
Chidi Happy Onyeoru

A thesis submitted to

The University of Sheffield



in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

16th March 2026

Chidi Happy Onyeoru

Measuring and Modeling of Human Trust in Machine using Brain Signals

PhD Thesis, 16th March 2026

Supervisors: Dr Mahnaz Arvaneh

The University of Sheffield

School of Electrical and Electronic Engineering

Amy Johnson Building

Portobello Street

Sheffield, S1 3JD

To Almighty God, and to my beloved wife, Josephine, and our children, Freda, Emmanuella, and Emmanuel.:

Chidi.

Acknowledgement

I would like to express my sincere gratitude to my family for their unconditional love, patience, and unwavering support throughout my PhD journey. I am deeply thankful to my supervisor, Professor Mahnaz Arvaneh, for her invaluable guidance, encouragement, and mentorship. Her support and belief in my potential have been instrumental in sustaining my motivation during this demanding yet rewarding academic journey. I also extend my heartfelt appreciation to all members of our research team for their support, collaboration, and encouragement.

Declaration

I, Chidi Happy Onyeoruru, hereby declare that the work presented in this thesis is entirely my own original research undertaken as part of my PhD studies. Any material that has been derived from the work of others has been appropriately acknowledged and referenced.

Sheffield, 16th March 2026

Chidi Happy Onyeoruru

Abstract

Human trust is a fundamental component of Human–Machine Interaction (HMI), especially in the context of autonomous systems. This study investigates the key determinants of trust in machines, focusing on machine error rates and types, as well as the degree to which humans can intervene and correct such errors. To explore these factors, three experimental scenarios were developed using a simulated claw robot tasked with navigating through multiple objects to identify and retrieve a designated target. The first experiment examined the effect of varying error rates on participants’ trust, the second investigated the influence of speed–accuracy trade-offs, and the third explored how the ability to intervene and rectify severe machine errors affected trust. A regression model was subsequently developed to estimate human trust levels based on machine performance characteristics. Results showed that error rate significantly influenced trust, with participants reporting higher trust when the robot operated at lower speed but with greater accuracy. Trust further increased when participants were allowed to intervene and correct the robot’s errors. Regression analyses confirmed that trust could be predicted based on different types of machine errors, offering insights for enhancing human–autonomy interaction in real-world applications. In addition, the study examined the neural correlates of trust by analyzing components of error-related potentials (ErrPs), including error-related negativity (ERN), error positivity (Pe), and P300. To assess the feasibility of real-time trust monitoring, a subject-specific Linear Discriminant Analysis (LDA) classifier was trained using selected ERP features. The classifier achieved an average accuracy of 79%, demonstrating the discriminative power of trust-related neural signatures. These findings underscore the potential of EEG-based adaptive systems for monitoring users’ trust levels and dynamically adjusting robot behavior in real time.

Contents

Chapter 1

Introduction to the Research Work1

- 1.1 Project Background 1
 - 1.1.1 Motivation 3
 - 1.1.2 Research Aim and Objectives 4
 - 1.1.3 Dissertation Outline 5

Chapter 2

Background Studies7

- 2.1 Introduction 7
 - 2.1.1 Concept of Trust in Human-Machine Interactions 7
 - 2.1.2 Measuring Human Trust Using Brain Signals 9
- 2.2 Brain-Computer Interfaces; Pathways for Brain Signal Generation and Trust Estimation 10
 - 2.2.1 Operational Paradigms of BCIs: Active, Reactive, and Passive Systems 13
- 2.3 Commonly Used Brain Imaging Techniques and Rationale for Electroencephalography (EEG) 15
 - 2.3.1 Components of EEG Signals 18
 - 2.3.2 Error-Related Potentials (ErrPs) in EEG Family and BCI Framework
20

2.3.3	Classification Algorithms Used in EEG Signal Classification . . .	22
2.3.4	Classifications of ErrP Responses	25
2.3.5	Commonly Used Classification Algorithms, and Extracted Features of Some Selected ErrP Studies	26
2.3.6	Relevance of ErrP Studies in BCI Systems Development	27
2.3.7	Methods for Different Applications of ErrP in BCI Real-Time System Control	29
2.4	Chapter Summary	30

Chapter 3

Analysing and Modeling of Human Trust to a Navigation Robot 33

3.1	Chapter Introduction	33
3.2	Methods	35
3.2.1	Participants	35
3.2.2	Experiment	36
3.2.3	Task	37
3.2.4	Error Definition	39
3.2.5	Data Analysis	40
3.3	Results	44
3.3.1	Scenario One	44
3.3.2	Scenario Two	45
3.3.3	Scenario Three	46
3.3.4	Estimating Human Trust Using Regression Analysis	46
3.3.5	Discussion	47
3.4	Chapter Summary	49

Chapter 4

Investigating Human Trust in Machine Using ErrP Signals . 51

4.1	Chapter Introduction	51
-----	--------------------------------	----

4.2	Methods	53
4.2.1	Participants	53
4.2.2	The Claw Task	54
4.2.3	EEG Data Acquisition and Experimental Setup	55
4.2.4	Preprocessing and Extraction of ErrP Components for Trust Analysis	56
4.2.5	Time-Domain Analysis of Error-related Potentials (ErrP) across Trust Conditions	57
4.2.6	Feature Extraction and Selection for Classification	59
4.2.7	Machine Learning-Based Classification of ErrP Features for Trust Level Prediction	60
4.3	Results	64
4.3.1	Results of Neurophysiological and Statistical Analysis	64
4.3.2	Classification Performance of LDA and SVM Models	66
4.3.3	Discussion	68
4.4	Chapter Summary	70

Chapter 5

Investigating Human Trust in Machine Using P300 Signals 73

5.1	Chapter Introduction	73
5.2	Methods	74
5.2.1	Participants	74
5.2.2	The Claw Task	75
5.2.3	EEG Acquisition and Experimental Setup	75
5.2.4	EEG Preprocessing and P300 Component Extraction for Trust Condition Analysis	76
5.2.5	P300 Component Identification and Statistical Analysis of Trust Conditions	77
5.2.6	Feature Extraction and Selection for Classification	77

5.2.7	Machine Learning Classification of P300 Features across Trust Conditions	79
5.3	Results	80
5.3.1	Neurophysiological and Statistical Analysis	81
5.3.2	Classification Using LDA and SVM	81
5.3.3	Discussions	82
5.4	Chapter Summary	84

Chapter 6

General Discussions, Future works and Conclusions87

6.1	General Discussions	87
6.2	Limitation	89
6.3	Future works	90
6.3.1	Expansion to Real-World and Multimodal HRI Environments ..	90
6.3.2	Longitudinal and Adaptive Trust Modeling	90
6.3.3	Cross-Individual and Cross-Context Generalization	91
6.3.4	Integration into Closed-Loop Trust-Aware Robotic Systems ...	91
6.3.5	Exploration of Additional Neural Markers and Cognitive States	91
6.3.6	Ethical and Societal Considerations in Neural Trust Monitoring	92
6.3.7	Multidisciplinary Integration and Cross-Disciplinary Collaboration	92
6.3.8	Summary	92
6.4	Conclusions	93
6.4.1	Behavioural Analysis of Human Trust Dynamics in Response to Robot Performance and Error Characteristics	93
6.4.2	Error-related Potentials as Electrophysiological Markers of Trust in Human-Robot Interaction	94
6.4.3	P300 Responses as Neurophysiological Indicators of Trust during Observation of Correct Robotic Actions	94
6.4.4	Summary	95

Introduction to the Research Work

1.1 Project Background

As intelligent machines become increasingly integrated into daily life and professional environments, understanding their behavior and intentions is essential for maintaining effective human control. A reliable framework for measuring and fostering trust is therefore critical to achieving a harmonious and efficient relationship between humans and machine agents. Prior research has shown that when users lack accurate knowledge of an automated system's performance, they often experience a mismatch between expectations and actual outcomes [1]. For instance, users of navigation applications may encounter difficulties when they do not fully understand or trust the system. During mobile network failures or signal delays, such users might repeatedly reset the application or deviate from the suggested route in an attempt to elicit a faster response. These repeated interventions can further disrupt system performance, rendering the application less efficient and reliable. Addressing and mitigating these expectation mismatches remains a fundamental challenge in the design and evaluation of autonomous and automated systems [1].

Effective human-machine collaboration requires maintaining an optimal balance of trust, wherein humans intervene only when necessary while allowing the system sufficient autonomy to perform its tasks. As noted by Jung et al. [1, 2], insufficient trust can lead to excessive hu-

man intervention even in well-functioning systems, whereas excessive trust may increase vulnerability to automation failures. The level of trust users place in a system shapes their overall perception of the agent and influences the quality of their interaction [2]. Previous studies have demonstrated that both human-related factors (e.g., experience, cognitive load) and agent-related factors (e.g., transparency, performance) contribute to trust under varying conditions [3]. Moreover, environmental aspects such as the operational context and task complexity can further affect the efficiency and performance of machine agents [4].

Beyond behavioral and environmental determinants, recent research has begun to explore the neural correlates of trust. Adolphs et al. [5] examined neural mechanisms underlying interpersonal trust, while other studies have employed electroencephalography (EEG) and galvanic skin response (GSR) to assess trust and distrust during simulated driving tasks [6]. Functional magnetic resonance imaging (fMRI) studies have similarly investigated differential brain activity in response to trustworthy versus untrustworthy human faces, as well as the neural mechanisms that support trust formation during human–human interaction [7, 8].

Although self-report measures remain the most common method for assessing human trust in machines, they present notable limitations. As highlighted by [1], continuous self-reporting interrupts the interaction process and may introduce bias or unreliability. Physiological measures such as GSR and heart rate variability (HRV) offer complementary insights but suffer from limited temporal precision. In contrast, EEG provides high temporal resolution on the order of milliseconds, making it a promising tool for continuously and objectively monitoring trust in real time.

Recent work has suggested associations between human confidence and error-related positivity (oPe) in EEG signals [1]; however, such studies have often been constrained by small sample sizes and limited experimental paradigms. Consequently, neurocognitive knowledge of trust in human–machine interaction remains limited, and there is a pressing need to identify robust EEG-based markers of trust [9].

This research seeks to advance the understanding of the neurological basis of human trust in machines by focusing on EEG correlates such as Error-related Potentials (ErrPs) and the P300 component [1, 2]. The first objective is to examine how variations in error probability, error severity, and the possibility of human intervention influence trust in autonomous systems. Furthermore, the study investigates whether specific ErrP components—namely, Error-related Negativity (ERN) and Error Positivity (Pe)—as well as the P300 component, correlate with dynamic changes in trust.

To address these objectives, participants engaged in interactive experiments involving a

simulated navigation robot. EEG signals were recorded as participants observed the robot's behavior, and subjective trust ratings were collected periodically via computer input. These trust ratings were then mapped onto EEG features to develop a classifier capable of detecting trust fluctuations based on neural markers.

The findings from the initial experiments are presented in Chapter 3, which models the effects of error-related factors and human intervention on trust. Chapters 4 and 5 build upon these results by analyzing EEG-derived markers, demonstrating how ERN, Pe, and P300 components can be utilized to model and predict human trust in real time. Collectively, these findings contribute to the development of adaptive, neurophysiologically informed frameworks for enhancing transparency, reliability, and cooperation in human-autonomy interaction.

1.1.1 Motivation

Human trust in machines remains insufficiently explored, despite its critical importance in human-machine collaboration. Previous studies have reported mismatches in automation performance resulting from both under-trust and over-trust in agent partners [10–12]. For effective and harmonious human-machine interaction, a certain degree of mutual understanding must be established, as this directly influences collaborative performance and the achievement of shared goals.

Although several studies have sought to examine trust in machines, further research is required due to the multifaceted nature of trust. Factors such as task complexity, machine operating conditions, algorithmic reliability, and user experience represent environmental and contextual variables that may undermine trust and, consequently, degrade system performance.

Motivated by recent advances in brain-computer interfaces (BCI), human-machine Interaction (HMI) technologies, this research aims to address the following gaps:

- The limited evidence linking error-related potential and P300 components to trust in human-machine interaction.
- The lack of real-time trust monitoring systems, and the use of non-realistic tasks in the context of manufacturing for trust measurement by previous studies.

To address these challenges, this study specifically examines how factors such as error probability, error severity, and the human ability to intervene in correcting machine errors

influence trust. Building on these findings, the study explores error-related potentials (ErrPs)—including error-related negativity (ERN) and error-positivity (Pe)—as potential neural indicators of trust, given prior evidence of their strong classification performance alongside the P300 component.

Ultimately, the results of this investigation provide a basis for the development of a BCI system capable of decoding human trust levels in real time. Such a system would enable continuous and accurate measurement of trust during human-machine interaction, thereby advancing the design of adaptive and reliable collaborative technologies.

1.1.2 Research Aim and Objectives

The primary aim of this study is to investigate the neurophysiological markers of human trust by analyzing electroencephalographic (EEG) responses to machine actions, with a particular focus on both correct and erroneous actions and their associated probabilities. The block flow diagram of the objectives is shown in Figure 1.1. By characterizing how trust is encoded in neural activity, this study seeks to advance the conceptual and practical understanding of trust in human-machine interactions.

A conceptual understanding of human trust in machines, anchored in measurable neurophysiological activity, provides a critical foundation for the design and implementation of brain-computer interface (BCI) systems. Such systems would enable continuous and reliable estimation of trust levels in real time, thereby supporting adaptive human-machine interactions. For instance, BCIs informed by trust-related EEG markers could regulate the degree of machine autonomy or adjust information transparency in response to users' fluctuating trust states, thus preventing overreliance or disuse.

To achieve this overarching aim, the study is guided by the following specific objectives:

- a) Investigate how error probability, error severity, and the ability to intervene influence human trust in the machine.
- b) Identify EEG responses to correct machine actions and examine how their components correlate with human trust.
- c) Identify EEG responses to erroneous machine actions and examine how their components correlate with human trust.
- d) Develop and implement machine learning methods to monitor human trust during

human-machine interactions.

In summary, this project is designed to bridge the gap between conceptual models of trust and their neurophysiological manifestations, with a view toward building real-time adaptive systems. By combining behavioral analysis, EEG-based correlates, and computational modeling, it advances both the theoretical understanding of trust and its practical applications in BCI-enabled human-machine collaboration.

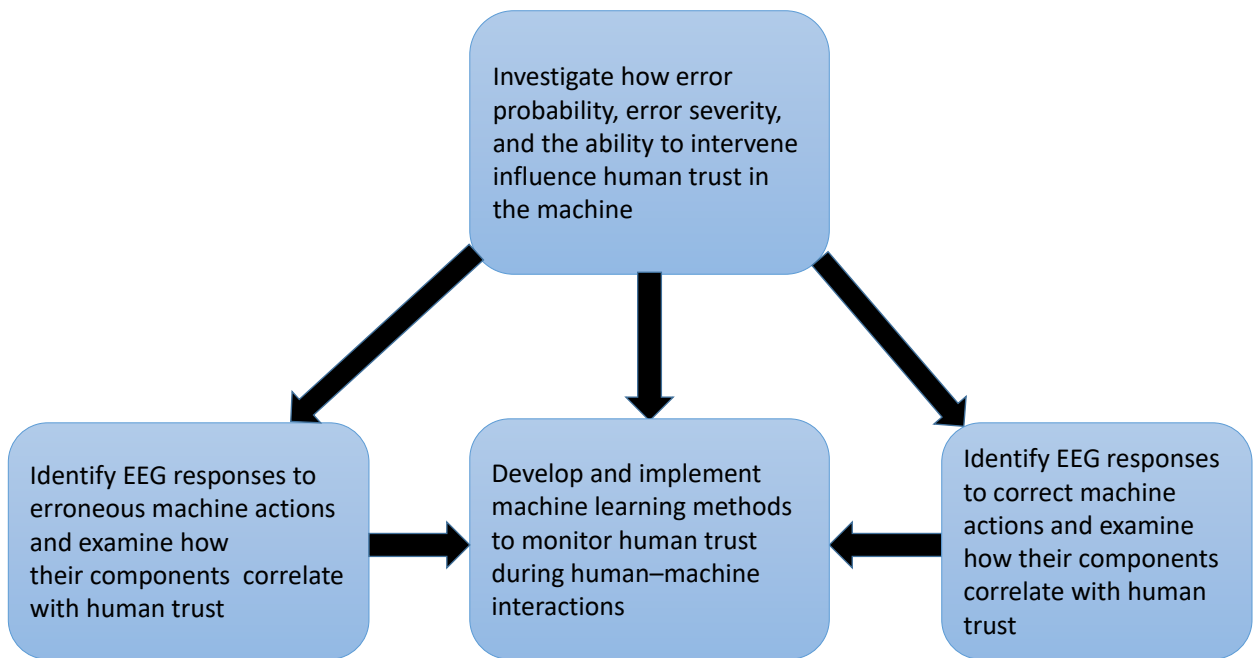


Figure 1.1: Block Diagram of Research Objectives.

1.1.3 Dissertation Outline

This dissertation is presented as:

- Chapter 1: Introduction to Research Work
- Chapter 2: Dissertation Background
- Chapter 3: Analysing and Modeling of Human Trust to a Navigation Robot
- Chapter 4: Investigating Human Trust in Machine Using ErrP Signals
- Chapter 5: Investigating Human Trust in Machine Using P300 Signals

- Chapter 6: Conclusion and Future Work

Background Studies

2.1 Introduction

2.1.1 Concept of Trust in Human-Machine Interactions

In general terms, trust can be defined as a psychological state involving the willingness of an individual to accept vulnerability based on positive expectations of the intentions, capabilities, and reliability of another party [13, 14]. When translated into measurable terms, trust can be operationalized through behavioral indicators (e.g., reliance on or compliance with automation recommendations), self-reported measures (e.g., questionnaires assessing perceived reliability or confidence), and physiological or neurophysiological markers (e.g., workload, error-related brain potentials, or hemodynamic responses). These measurable dimensions allow researchers to quantify trust, despite its abstract psychological foundation.

In human–human interactions, social and psychological behaviors—such as cooperation, honesty, and predictability—have traditionally been employed as indicators of trust. These approaches have been extended to human–machine interactions, where factors such as perceived competence, transparency, and predictability of machine agents serve as proxies for trust [1, 6]. However, reliance solely on social and psychological cues has not yielded con-

sistently reliable measures in the context of automation. This inconsistency is largely due to the unique dynamics of human-machine systems, where machines lack genuine social intention but are evaluated as if they do [1, 2].

Trust in automation is particularly critical because it underpins user confidence and determines the extent to which humans are willing to rely on machine partners for achieving common objectives. Without sufficient trust, users may underutilize reliable automation; conversely, excessive trust can result in overreliance and unsafe outcomes [4, 6, 15]. Consequently, substantial research has focused on identifying measurable determinants of trust and clarifying the contexts in which they should be applied [1, 2, 6, 13].

Bisantz and Seong [4] investigated key determinants of trust, identifying functional reliability of the machine, the quality of information or decision support provided, and the handling of system failures as central factors. They emphasized that trust is often eroded by failures arising from environmental disturbances, flaws in automation algorithms, poor decision-support mechanisms, or limitations in the human-machine interface. Such anomalies contribute to unpredictability in system behavior, which undermines both user trust and overall system utilization.

Other studies have focused on real-time and dynamic measurement approaches. Hu et al. [6] demonstrated that psychological signals can be mapped in real time to estimate human trust levels, thus enabling adaptive human-machine systems that respond to fluctuating trust states. Complementary work by Aiken et al. [16, 17] showed that a user's awareness of a robot's limitations—rather than its intentions—enhances cooperative behavior and sustains confidence in the system. In parallel, Hancock et al. [18] identified reliability and error rates as quantifiable determinants of trust in automation. Similarly, Salem et al. [19] examined the social consequences of robotic errors, reporting that user cooperation decreases following errors but may be partially restored if the robot acknowledges mistakes or issues an apology [16]. These findings highlight that trust is not only influenced by technical reliability but also by the machine's perceived social behavior and accountability.

Building on these empirical findings, Hoff and Bashir [14] proposed a comprehensive framework categorizing trust into three components: dispositional, situational, and learned. Dispositional trust reflects stable, trait-like factors such as demographic attributes (e.g., gender, culture) and general attitudes toward technology. Situational trust depends on contextual conditions, including task type, difficulty, and environmental uncertainty. Learned trust evolves over time based on prior experiences with the machine and its observed performance. Of these, situational and learned trust are highly dynamic and exert significant influence on

real-time human decision-making during interaction with machines. This framework underscores the need for multidimensional approaches to measuring trust, integrating subjective, behavioral, and physiological indicators to capture its complexity.

In summary, trust in human-machine interaction is best conceptualized as a dynamic, measurable construct that integrates psychological expectations, behavioral reliance, and physiological responses. Effective measurement requires a multimodal approach that not only accounts for static factors such as demographics but also tracks dynamic influences such as workload, system reliability, and real-time neurophysiological signals. Establishing such measures is essential for designing adaptive automation systems that can calibrate user trust appropriately, thereby enhancing safety, efficiency, and collaboration.

2.1.2 Measuring Human Trust Using Brain Signals

As stated above, the measurement of human trust requires approaches that differ across disciplines and application domains. Trust is shaped by multiple parameters, including contextual conditions, cognitive states, emotional responses, and social cues, which makes its assessment particularly challenging in dynamic human-machine interaction environments [3]. While self-report measures and behavioral observation have traditionally been employed, recent research has emphasized the need for neurophysiological methods to provide objective, real-time insights into the underlying mechanisms of trust.

In human-robot interaction, cognitive workload has been identified as a key factor influencing fluctuations in trust [20–29]. Hirshfield et al. provided one of the seminal contributions in this area by combining psychological testing with neuroimaging methods such as electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) to investigate emotional and cognitive correlates of trust [20]. Their study focused on the dynamic transition between trust and distrust, particularly in scenarios where participants detected deception by a computer agent. The authors proposed that emotional-cognitive states such as workload, frustration, and surprise are directly linked to decreases in trust. Neurophysiological data supported this hypothesis, as reduced trust was associated with increased workload and frustration, reflected in measurable changes in oxygenated and deoxygenated hemoglobin concentrations within both cerebral hemispheres. This work not only demonstrated the feasibility of using brain signals to assess trust but also highlighted the intertwined nature of cognitive load, emotional state, and trust dynamics.

Other studies have examined anthropomorphism as a determinant of trust. By diversifying

agents with internal and external human-like features, researchers sought to understand the degree to which human resemblance affects trust in non-reciprocal interactions [2, 13]. Using EEG, they recorded neural responses to variations in trust levels and found that agents with stronger external human-like features elicited heightened brain activity correlating with trust [2, 13]. Similarly, neuroimaging studies have demonstrated that fMRI can identify brain regions activated in response to trustworthy versus untrustworthy human faces [7, 8, 30, 31].

EEG and fNIRS offer unique advantages for studying trust by providing fine-grained temporal and physiological data on how the brain responds to changes in interaction context. The integration of these techniques has the potential to reveal multi-dimensional markers of trust, capturing both rapid neural responses and slower hemodynamic changes. However, the application of these tools to human-machine interaction scenarios is still underexplored, with most existing studies conducted in controlled laboratory environments that may not fully capture the complexity of real-world interactions.

Despite the growing interest in this field, neurocognitive understanding of trust in human-machine interaction remains limited and underdeveloped. Although researchers increasingly recognize the importance of trust as a key determinant of effective collaboration with automated systems, the underlying neural mechanisms that shape trust formation, maintenance, and breakdown are still not well understood. To date, only a small number of studies, for instance [1, 2], have directly investigated neurophysiological markers specifically associated with trust, leaving substantial gaps in the literature. Furthermore, the existing findings are often based on narrow experimental conditions or small participant samples, which limits their generalizability. To advance the field, additional research is required to systematically examine how trust relates to a broader range of EEG components, how these neural signatures vary across contexts and individual differences, and how they evolve dynamically as human users interact with machines over time.

2.2 Brain-Computer Interfaces; Pathways for Brain Signal Generation and Trust Estimation

Brain-Computer Interfaces (BCIs) provide a non-muscular communication pathway between the human (or animal) brain and external devices, such as computer systems, assistive technologies, and robotic platforms [32]. The growing demand for portable, efficient, cost-effective, and reliable technologies that can harness brain activity for monitoring and control has fueled extensive research in this area. Among the available approaches, non-invasive electroencephalography (EEG) has become the most widely adopted due to its portability,

safety, and high temporal resolution. EEG-based methods will be discussed in greater detail in subsequent sections of this report.

Current BCI systems allow users to control external devices—such as wheelchairs, prosthetic limbs, and computer cursors—by decoding EEG signals generated during motor imagery (MI) tasks, in which imagined movements of body parts are translated into commands [9, 33]. Beyond MI, other paradigms such as Steady-State Visual Evoked Potentials (SS-VEPs) and P300 event-related potentials expand the scope of BCIs by leveraging visually or cognitively evoked brain responses to stimuli, enabling flexible and individualized modes of interaction [34, 35].

A particularly promising line of research involves the use of Error-Related Potentials (ErrPs) as control signals within BCIs. ErrPs are neural responses spontaneously generated when individuals detect an error—whether their own, another human’s, or a machine’s—while observing actions through an interface [36]. These responses can be reliably captured using EEG and subsequently classified with machine learning techniques [37–40]. By doing so, BCIs are able to infer whether a user perceives an action as correct or erroneous, thus creating a feedback mechanism for adaptive system behavior [41, 42].

Incorporating ErrPs into BCI frameworks moves machines toward semi-autonomous operation, where system decisions are continually refined to align with user expectations. This reduces the demand on users for constant low-level control and instead positions them as high-level supervisors. Importantly, this mechanism also provides a unique opportunity to study and evaluate human trust in machines. Since ErrPs capture real-time neural responses to perceived errors, they offer a direct and objective measure of user confidence in machine actions. For instance, frequent ErrP responses to a system’s decisions may indicate declining trust, whereas reduced ErrP activity in the face of accurate and reliable performance may suggest strengthening trust.

By integrating ErrP detection with adaptive machine learning models, BCIs can serve as both control systems and trust evaluation tools. This dual role enables the development of intelligent human-machine systems that dynamically adjust their behavior to maintain an optimal balance between autonomy and user oversight. In this way, BCIs not only enhance functional interaction but also open new avenues for systematically studying and modeling the dynamics of human trust in machines—an increasingly critical factor in the deployment of autonomous and assistive technologies.

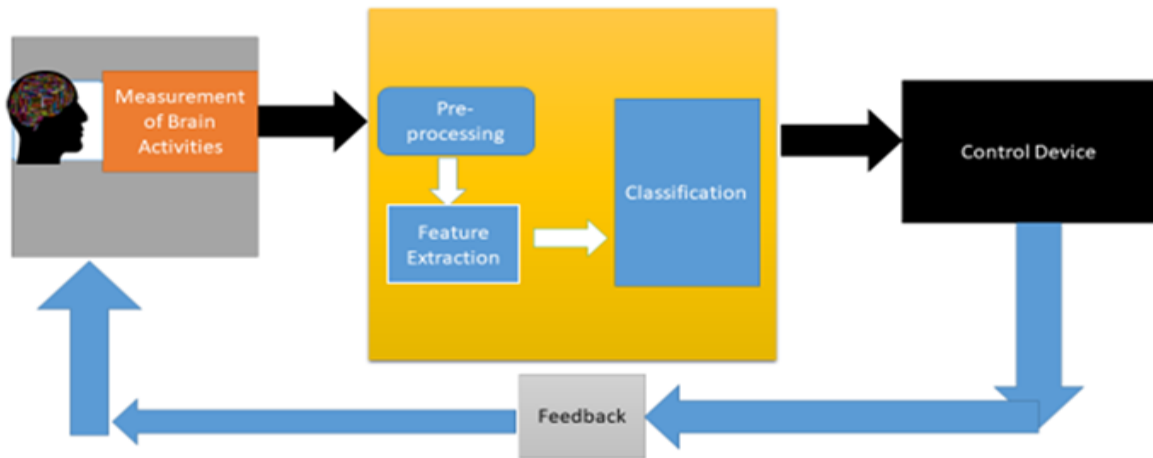


Figure 2.1: Schematic Representation of a Simple Brain-Computer Interface (BCI) System

Fundamental Components of a Basic Brain-Computer Interface (BCI) System : A Brain-Computer Interface (BCI) system is generally comprised of four fundamental components: (i) brain signal acquisition, (ii) signal processing, (iii) a control device (output interface), and (iv) feedback. Each of these components plays a vital role in ensuring the effective translation of neural activity into executable commands for external devices. A simplified block diagram of a basic BCI system is presented in Figure 2.1.

- **Data Acquisition Unit:** The data acquisition stage represents the first and one of the most critical parts of the BCI system. It involves the recording of neural activity using electrodes or other electrical sensors placed on the scalp of a human or animal subject. The acquired brain signals represent the user's cognitive or motor intentions, which are subsequently processed for control purposes. Importantly, this unit also functions bidirectionally, receiving feedback from external devices, thereby closing the interaction loop between the user and the system.
- **Signal Processing Unit:** The signal processing module comprises three key sub-units:
 - **Pre-processing:** The pre-processing stage refines the raw EEG or other neural signals by removing noise and artifacts (e.g., ocular or muscular activity), thereby improving the signal-to-noise ratio. The pre-processed data are then forwarded to subsequent stages for further analysis.
 - **Feature Extraction:** In this stage, salient information is extracted from the pre-processed neural activity and converted into representative feature val-

ues. These features serve as compact descriptors of brain activity and form the basis for generating machine-readable control commands [32]. - Classification: The classification stage discriminates between relevant neural features (e.g., those corresponding to a specific imagined or perceived action) and unrelated features. By identifying task-specific patterns, the classifier translates user intent into a discrete output that can be used to control external devices [9, 32, 33].

- **Control Device (Output Interface):** The control device serves as the output stage of the BCI system. It receives encoded neural commands from the classifier and executes them in real time. Depending on the application, this device may take the form of a computer system, a wheelchair, or a robotic manipulator. Through this interface, the user's cognitive intentions are translated into tangible physical actions [9, 32].
- **Feedback Module:** Feedback represents the final stage of a closed-loop BCI system, providing users with information about the outcomes of their neural commands. Feedback modalities may include visual, auditory, or tactile cues, each of which enables the user to monitor system performance and adapt their mental strategies accordingly. By reinforcing the link between intention and outcome, feedback significantly enhances learning, control accuracy, and overall system performance [32].

Overall, the integration of these four components—signal acquisition, signal processing, control device, and feedback—forms the backbone of BCI systems. Together, they enable the real-time translation of brain activity into meaningful interactions with external devices, thereby laying the groundwork for practical applications in rehabilitation, assistive technologies, and human-machine collaboration.

2.2.1 Operational Paradigms of BCIs: Active, Reactive, and Passive Systems

The operational paradigms of Brain-Computer Interfaces (BCIs) have evolved over time to accommodate different forms of human-machine interaction. Early BCI research primarily focused on active BCIs, where the user intentionally modulates brain activity to control external devices. Over the years, however, the classification of BCIs has expanded to include passive and reactive systems, depending on the nature of the neural signals utilized and the role of user intention in generating these signals.

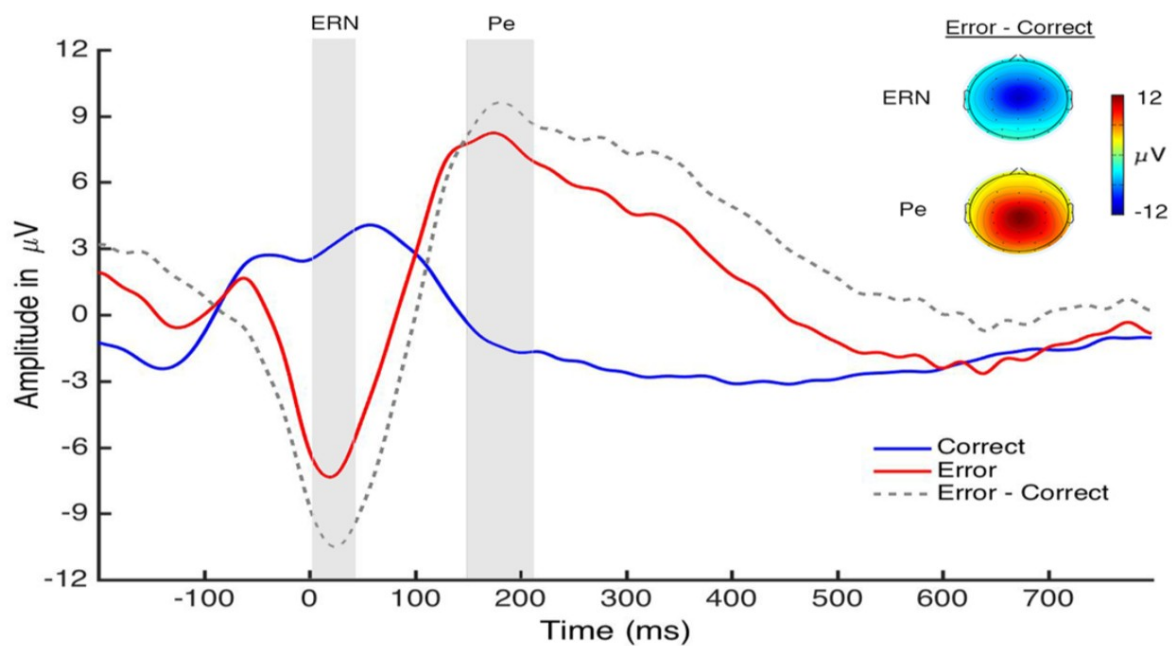


Figure 2.2: Performance event-related potentials (ERPs): Components of ERN and Pe. The graph illustrates the response-locked error-related negativity (ERN) and error positivity (Pe) elicited during task performance. The gray boxes indicate the analysis windows used to quantify each component [43]

Active BCIs. Active BCIs are systems in which users deliberately generate brain responses to convey specific intentions for external device control. These paradigms are predominantly based on sensorimotor rhythms (SMRs), such as those employed in motor imagery (MI)-based BCI systems, where imagining movements of specific body parts (e.g., left or right hand) produces detectable EEG patterns that can be decoded into commands [33]. Active BCIs also include systems driven by slow cortical potentials (SCPs), which represent longer-term shifts in cortical activity and have been applied in communication and assistive technologies. These paradigms remain a cornerstone of BCI research, as they directly translate intentional brain activity into meaningful system outputs.

Passive BCIs. In contrast, passive BCIs are designed to derive outputs from spontaneous or involuntary brain activity, without requiring the user to consciously control the system. Zander and colleagues first formalized this paradigm, emphasizing that passive BCIs operate independently of volitional commands from the user [44, 45]. A practical example of a passive BCI is a system that continuously monitors the user's mental workload and adaptively reduces distractions or task demands to optimize performance [46–48].

Error-related potentials (ErrPs) provide another example of passive BCIs. ErrPs are automatically elicited when a user detects an error, either in their own action or in the behavior

of a machine. These signals can be captured via EEG and processed to improve system reliability by autonomously correcting errors, without requiring conscious user intervention [47, 49].

Reactive BCIs. A third category, termed reactive BCIs, was later introduced by Zander and Kothe to account for systems in which neural outputs are elicited by responses to external stimuli rather than spontaneous or intentional activity [46, 47]. Reactive BCIs leverage the brain's natural reactions to presented stimuli, such as steady-state visual evoked potentials (SSVEPs), which occur when a visual stimulus flashes at a fixed frequency, leading to EEG signals at the same frequency [50]. Another widely studied reactive paradigm is the P300 response, an event-related potential elicited when a rare or significant stimulus is presented among a sequence of irrelevant stimuli, as described earlier in this report.

Interestingly, based on this classification framework, ErrP-based BCIs can also be considered reactive, since these potentials are generated specifically in response to erroneous actions performed by either the user or the machine. Thus, ErrP paradigms occupy a unique position at the interface of passive and reactive BCIs, depending on whether they are interpreted as spontaneous error monitoring or as stimulus-driven responses to observed mistakes.

Finally, the categorization of BCIs into active, passive, and reactive systems reflects the growing diversity of paradigms that exploit different aspects of human neurophysiology. Active BCIs rely on deliberate user control, passive BCIs harness spontaneous or involuntary brain activity, and reactive BCIs exploit stimulus-driven neural responses. This classification not only provides a theoretical framework for understanding BCI systems but also highlights the wide range of applications—from intentional device control to adaptive error correction and context-aware support—that BCIs can enable.

2.3 Commonly Used Brain Imaging Techniques and Rationale for Electroencephalography (EEG)

Over the years, a variety of neuroimaging techniques have been developed and applied to the measurement of brain activity in both research and applied contexts. These methods include Electroencephalography (EEG), Magnetoencephalography (MEG), Functional Magnetic Resonance Imaging (fMRI), Positron Emission Tomography (PET), and Near-Infrared Spectroscopy (NIRS). Each of these approaches provides unique insights into neural functioning, but

they also differ in terms of temporal resolution, spatial resolution, portability, invasiveness, and cost. As a result, the choice of technique is often dictated by the research objectives, experimental setting, and practical constraints.

Magnetoencephalography (MEG) measures magnetic fields generated by neuronal electrical activity, offering excellent temporal resolution and reasonable spatial localization. However, MEG systems are costly, require magnetically shielded rooms, and lack portability, which limits their application outside of controlled laboratory environments.

Functional Magnetic Resonance Imaging (fMRI) measures blood oxygenation level-dependent (BOLD) signals, thereby providing high spatial resolution for mapping brain activity across cortical and subcortical regions. Despite these advantages, fMRI suffers from poor temporal resolution, is highly sensitive to motion artifacts, and requires large, immobile, and expensive scanners, making it unsuitable for real-time or portable applications.

Positron Emission Tomography (PET) employs radioactive tracers to quantify metabolic activity in the brain. While PET provides valuable insights into long-term neurochemical processes, it is invasive, has low temporal resolution, and involves significant safety considerations due to radiation exposure, restricting its use for continuous or repeated measurements.

Near-Infrared Spectroscopy (NIRS) estimates brain activity by measuring changes in oxygenated and deoxygenated hemoglobin in cortical regions. NIRS is portable, relatively inexpensive, and non-invasive, but it is limited by shallow penetration depth and lower temporal resolution compared to electrophysiological techniques. The comparative pros and cons of these techniques are shown in Table 2.1

Among these methods, Electroencephalography (EEG) has emerged as the most practical and widely used technique for measuring brain activity in applied contexts, particularly within human-machine interaction studies [32]. EEG records electrical activity generated by synchronized postsynaptic potentials in cortical neurons, offering several distinct advantages:

- High temporal resolution: EEG captures brain activity on a millisecond timescale, enabling the study of rapid neural responses during dynamic interactions.
- Portability: Modern EEG systems can be configured in mobile or wearable formats, making them suitable for experiments outside traditional laboratory environments.
- Cost-effectiveness: Compared to MEG or fMRI, EEG equipment is significantly less expensive to acquire and maintain.

- **Compatibility:** EEG can be integrated with other measurement modalities (e.g., eye tracking, physiological sensors) for multimodal studies of cognition and behavior.
- **Non-invasiveness and safety:** EEG requires no exposure to radiation or strong magnetic fields, allowing repeated use without health risks.

While EEG has relatively lower spatial resolution compared to fMRI or MEG and is sensitive to artifacts from muscle activity or environmental noise, its advantages in portability, affordability, and temporal precision make it especially suitable for real-time monitoring of cognitive states such as trust, workload, and attention. Consequently, EEG is considered one of the most effective brain imaging tools for research on human-machine interaction and brain-computer interface (BCI) applications.

Despite its wide applicability and non-invasive nature, Electroencephalography (EEG) is not without limitations. Two of the most prominent drawbacks of EEG are its inherently low signal-to-noise ratio (SNR) and poor spatial resolution when compared to other neuroimaging techniques [32]. These challenges arise primarily from the fact that EEG signals are weak cortical activities that must travel through the skull and scalp before being detected, thereby making them highly susceptible to contamination from noise and artifacts.

To enhance the reliability and accuracy of EEG-based measurements, extensive efforts have been devoted to the development of signal preprocessing techniques. These methods are designed to minimize the presence of unwanted contaminants and improve the extraction of relevant neurophysiological information. Several well-established algorithms have been employed for this purpose, including:

- **Common Average Referencing (CAR)**, which reduces global noise by referencing each channel to the average of all channels;
- **Independent Component Analysis (ICA)**, which separates mixed signals into independent sources, often used for eye-blink and muscle artifact removal;
- **Common Spatial Patterns (CSP)**, a technique widely used in classification tasks to enhance discriminative features;
- **Principal Component Analysis (PCA)** and **Singular Value Decomposition (SVD)**, both of which reduce dimensionality while retaining the most significant variance in the data;
- **Surface Laplacian (SL)**, which improves spatial resolution by emphasizing local potentials; and

- Hybrid approaches, which combine two or more of the above methods for enhanced robustness [51].

Because EEG signals are predominantly slow-wave in nature, low-level impedance differential amplifiers are commonly employed during acquisition to ensure the accurate capture of low-frequency brain activities [32]. Additionally, to suppress recording artifacts, filtering techniques are widely implemented. A 50 Hz or 60 Hz notch filter is typically used to eliminate power-line interference, while high-pass filters remove low-frequency drifts, often caused by amplifier instabilities, thus ensuring that the extracted features reflect true neurophysiological activity [32].

In addition to algorithmic and hardware solutions, a number of practical considerations must be observed to ensure accurate EEG recordings. The positioning of electrodes using the internationally recognized 10–20 placement system remains a fundamental step for consistency and reproducibility [32]. Furthermore, special attention must be paid to the minimization of ocular artifacts through electrooculogram (EOG) monitoring and correction. In some experimental setups, subjects are required to stabilize their head position—for example, using a chin rest—to minimize motion artifacts. These measures, taken together, form an essential part of the precautionary framework that ensures EEG signals are as reliable and artifact-free as possible before subsequent preprocessing and neurophysiological analysis.

Table 2.1: Prons and Cons of Brain-Signal Measurement Techniques and Their Associated Neural Activities

Methods	Measurement Quality	Spatial Resolution	Temporal Resolution	Portability
EEG	<i>Electric Potential</i>	<i>Very Low</i>	<i>Very High</i>	<i>Good</i>
MEG	<i>Magnetic Potential</i>	<i>High</i>	<i>Very High</i>	<i>Not Good</i>
fMRI	<i>Hemodynamic Activity</i>	<i>Very High</i>	<i>Low</i>	<i>Not Good</i>
NIRS	<i>Hemodynamic Activity</i>	<i>Low</i>	<i>Low</i>	<i>Good</i>

2.3.1 Components of EEG Signals

An electroencephalogram (EEG) signal, in its raw form, represents a complex composite of multiple overlapping physiological and non-physiological sources. Unlike an idealized representation of brain activity, the scalp-recorded EEG is not inherently “clean.” Instead, it contains a superposition of meaningful neural information and various interfering signals. This is because EEG electrodes, placed non-invasively on the scalp, are highly sensitive not only to

Table 2.2: Classification of EEG Signal Components: Neural vs. Non-Neural

Category	Subcomponent	Description / Functional Significance
Neural Components	Brain Rhythms (Oscillations)	
	Delta (0.5–4 Hz)	Associated with deep sleep and unconscious states.
	Theta (4–8 Hz)	Linked with drowsiness, memory encoding, and emotional processes.
	Alpha (8–13 Hz)	Prominent during relaxed wakefulness; strongest in the occipital region with eyes closed.
	Beta (13–30 Hz)	Associated with active thinking, alertness, motor activity, and anxiety.
	Gamma (>30 Hz)	Linked to higher cognitive functions such as perception, attention, and memory binding.
	Event-Related Potentials (ERPs)	
	P300	Time-locked response to infrequent or significant stimuli.
	Error-Related Potentials (ErrPs)	Elicited when an error is perceived.
	Steady-State Evoked Potentials (SSEPs)	Responses to periodic sensory stimulation, e.g., SSVEPs in vision.
	Background Activity	Continuous spontaneous neural firing reflecting overall cortical state.
Artifacts and Noise	EMG Activity	Muscle movements (jaw clenching, forehead tension, blinking).
	EOG Activity	Eye movements and blinks, producing large slow waves.
	ECG Activity	Heartbeats visible in EEG, especially near temporal electrodes.
	Power Line Interference	50/60 Hz noise from electrical mains supply.
	Movement Artifacts	Electrode shifts, head movements, or cable vibrations.
	Environmental Noise	External disturbances from recording equipment or nearby electronics.

cortical neuronal activity but also to a wide range of additional biological processes and environmental disturbances. Consequently, the raw EEG must be regarded as a mixture of signals rather than a direct, uncontaminated measure of brain function.

For analytical clarity, these contributing components are typically categorized into two broad classes: neural signal components and non-neural signal components. Neural components originate from synchronous neuronal firing within the brain and are of primary interest in neurophysiological and clinical research. They reflect oscillatory brain rhythms, event-related potentials, and other patterns of cortical dynamics that are closely linked to sensory, cognitive, and motor processes.

In contrast, non-neural components—commonly referred to as artifacts and noise—do not originate from the brain but instead arise from either physiological activities outside the brain or from technical and environmental sources. Physiological artifacts include signals generated by ocular movements and blinks (electrooculographic activity), muscle contractions (electromyographic activity), and cardiac rhythms (electrocardiographic activity). Non-physiological noise, on the other hand, typically results from factors such as fluctuations in electrode impedance, power-line interference, or other recording system instabilities. These extraneous components can severely mask or distort the neural activity of interest, thereby complicating interpretation and analysis.

A structured classification of these EEG signal constituents, distinguishing between neural and non-neural sources, is presented in Table 2.2. Such a classification provides a conceptual framework for understanding the composition of raw EEG recordings and underscores the importance of preprocessing techniques aimed at artifact detection and removal.

2.3.2 Error-Related Potentials (ErrPs) in EEG Family and BCI Framework

One of the persistent challenges in Brain–Computer Interface (BCI) research is the inability of current systems to consistently acquire accurate neural signals or communication commands that fully reflect the user’s intended actions. This limitation arises largely from the susceptibility of BCIs to errors, which may originate either from the user (e.g., incorrect responses) or from the machine (e.g., misclassifications during calibration or execution). Such errors undermine the reliability of BCIs and hinder their widespread practical adoption.

To address this issue, research has increasingly focused on the exploitation of Error-Related

Potentials (ErrPs)—a class of event-related potentials (ERPs) that occur when an individual perceives an error in their own actions or in those of an external agent such as a computer system. ErrPs typically include three principal components: Error-Related Negativity (ERN or Ne), Late Error Positivity (Pe), and Pre-movement Positivity (PNe), each with distinct temporal and spatial characteristics [36, 52, 53].

These components are generated spontaneously when errors are detected, usually in the context of visual or auditory stimuli presented on a computer interface. Within BCI systems, ErrPs are of particular importance because they can be translated into digital commands using regression and classification algorithms. This enables real-time error detection and correction, thereby significantly improving system performance and reliability. ErrP-based BCIs thus represent a promising avenue for achieving more autonomous and adaptive human-machine interaction.

Error-related Negativity (ERN or Ne) The Error-Related Negativity (ERN or Ne) is a sharp negative deflection that appears approximately 80–150 ms after an erroneous motor response, or 40–80 ms after EEG calibration events [37, 38, 54]. Historically, the earliest description of ERN is attributed to Petrovna Bekhterev in 1968, who referred to it as an “error detector” [55]. Later, in 1990, two independent research groups led by Michael Falkenstein and W. J. “Bill” Gehring formally identified and named the component (Ne) while studying simple choice response tasks in which participants made mistakes [55].

The ERN is most prominent over fronto-central scalp sites, with maximal amplitudes typically recorded at electrodes in these regions. Researchers such as Spüler and Niethammer demonstrated that the ERN can be reliably identified using peak-to-peak amplitude differences, measured between the most negative peak occurring within 1–150 ms after response onset and the preceding positive peaks occurring 0–100 ms before the response [53]. The ERN is therefore regarded as a robust neural marker of error monitoring processes, playing a key role in adaptive behavior.

Late Error Positivity (Pe) Following the ERN, a Late Error Positivity (Pe) is typically observed. The Pe is characterized by a broad positive deflection with a centro-parietal distribution and occurs approximately 200–500 ms after the onset of an erroneous response [36, 52, 53]. The amplitude of the Pe has been shown to depend on error awareness, making it functionally analogous to the P300 wave, which is associated with conscious perception of stimuli [56, 57]. Thus, while the ERN is thought to reflect automatic error detection processes, the Pe represents a higher-level evaluation stage that involves conscious recognition of errors. Its

detection provides a secondary neural signal that can be exploited in BCIs for more reliable error classification.

Pre-Movement Positivity (PNe) Another important ErrP component is the Pre-Movement Positivity (PNe), which appears as a small positive voltage deflection occurring between –200 ms and –50 ms prior to the erroneous motor response [55]. This deflection is often considered a preparatory “excitation signal” that primes motor neurons for sensorimotor activity [58, 59].

Studies have shown that the amplitude of the PNe is typically smaller before erroneous responses compared to correct responses, suggesting its potential utility as an early neural marker for distinguishing between correct and incorrect actions [55, 59]. Furthermore, reduced PNe amplitudes have been reported in individuals who commit more frequent errors during Flanker tasks. This has clinical implications, as PNe could serve as a diagnostic or monitoring tool in populations prone to high error rates, such as individuals with Attention Deficit Hyperactivity Disorder (ADHD) or Obsessive–Compulsive Disorder (OCD) [55]. In summary, ErrPs represent a powerful neurophysiological signal family for enhancing BCI performance. Their components—ERN, Pe, and PNe—capture different stages of error processing, ranging from automatic detection to conscious recognition and preparatory activity. By integrating these signals into machine learning algorithms, BCIs can detect and correct errors in real time, paving the way for more robust and adaptive human–machine collaboration.

2.3.3 Classification Algorithms Used in EEG Signal Classification

Classification of electroencephalogram (EEG) signals is a central task in brain–computer interface (BCI) research and related biomedical applications. Due to the inherently noisy, high-dimensional, and non-stationary nature of EEG data, a variety of algorithms have been developed and applied, each with unique strengths and limitations. These algorithms can be broadly categorized into linear methods, kernel-based methods, tree-based approaches, instance-based learners, probabilistic models, and deep learning architectures.

Linear classifiers such as Linear Discriminant Analysis (LDA) and Logistic Regression are among the simplest and most widely used methods. LDA projects data into a space that maximizes class separation while minimizing within-class variance, making it computationally efficient and robust for low-dimensional features such as common spatial patterns (CSP). Logistic Regression, on the other hand, models the probability of class membership using a sigmoid function and is effective for binary EEG classification problems like seizure detection.

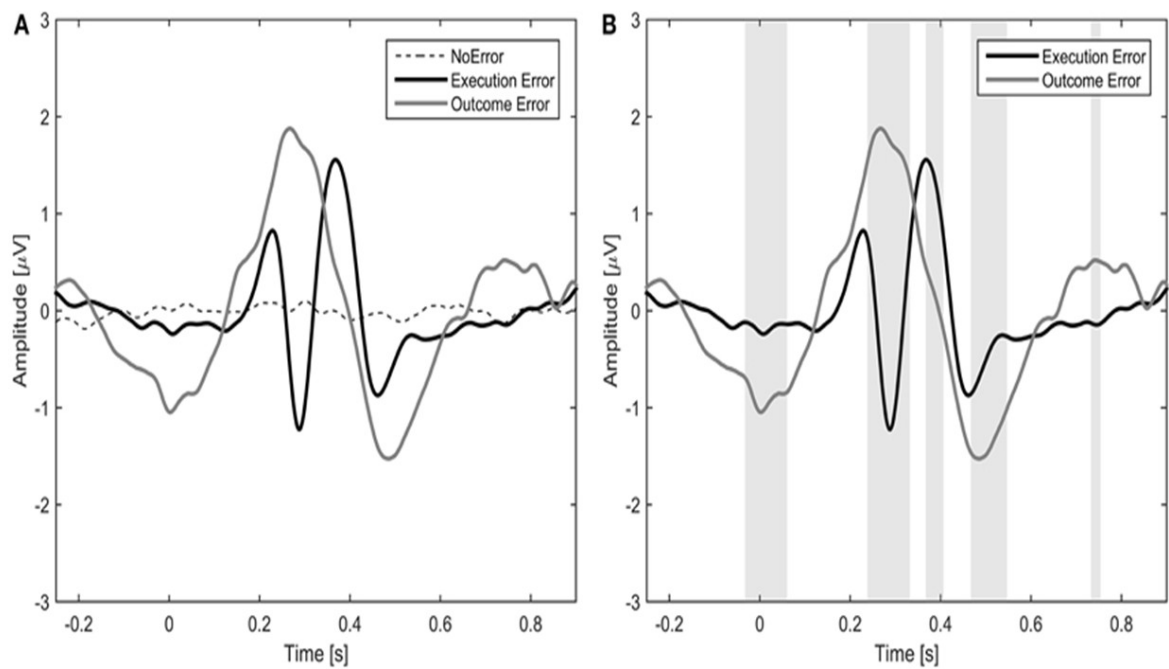


Figure 2.3: This figure illustrates the event-related potentials associated with execution and outcome errors. (A) Event-related potentials recorded at electrode FCz averaged across participants for NoError trials, execution errors, and outcome errors. (B) Comparison of event-related potentials for execution and outcome errors. The gray shaded regions indicate time intervals showing significant differences between execution and outcome errors ($p < 0.05$, Bonferroni corrected) [53].

Support Vector Machines (SVMs) extend linear classification by maximizing the margin between classes and, through the use of kernel functions, can capture complex non-linear boundaries in EEG data. They are particularly effective for high-dimensional EEG features, such as those derived from motor imagery tasks.

Tree-based methods including Decision Trees, Random Forests, and Gradient Boosting Machines (GBMs) are attractive due to their interpretability and flexibility. Decision Trees provide simple if-then rules but are prone to overfitting, whereas Random Forests and GBMs improve robustness by aggregating ensembles of trees. These ensemble approaches are capable of handling noisy EEG features and capturing non-linear relationships.

Instance-based classifiers such as k-Nearest Neighbors (k-NN) classify samples based on similarity to training data. Although conceptually simple and effective for well-separated EEG features, k-NN can be computationally expensive for large datasets.

Probabilistic classifiers such as Naïve Bayes and Hidden Markov Models (HMMs) exploit probabilistic assumptions to model EEG data. Naïve Bayes is fast and effective for certain EEG features, such as frequency band powers, but relies on the unrealistic assumption of feature independence. HMMs, in contrast, are particularly suited to temporal EEG data, for example in sleep stage classification, as they capture transitions between latent states.

Neural network-based methods have gained significant traction with the availability of larger EEG datasets and increased computational power. Multilayer Perceptrons (MLPs) are capable of modeling non-linear relationships but require sufficient training data. Convolutional Neural Networks (CNNs) have been successfully applied to raw EEG and time-frequency representations, learning spatial-temporal patterns directly from the data. Recurrent Neural Networks (RNNs), including LSTMs and GRUs, are effective for modeling sequential dependencies in EEG time series, such as seizure prediction. Hybrid CNN-RNN models combine spatial and temporal feature extraction, yielding strong performance in multi-class mental state classification.

Finally, specialized methods such as Extreme Learning Machines (ELMs), Gaussian Mixture Models (GMMs), and Quadratic Discriminant Analysis (QDA) have been explored in specific EEG applications. These approaches offer advantages in speed (ELM), soft classification (GMM), or non-linear decision boundaries (QDA), depending on the task requirements.

The choice of algorithm is therefore highly dependent on the characteristics of the EEG data, the task complexity, and practical constraints such as the need for real-time operation. Table 2.2 provides a comparative summary of commonly used EEG classification algorithms,

highlighting their underlying principles, strengths, limitations, and typical applications.

2.3.4 Classifications of ErrP Responses

Early investigations into error-related potentials (ErrPs) were primarily concerned with distinguishing between neural responses to correct and erroneous actions. This binary classification provided an important foundation for understanding how the brain monitors performance and detects mistakes. However, subsequent research has demonstrated that the neural signatures of error processing are far more nuanced, leading to the development of more refined classification frameworks. Only a limited number of studies have attempted to categorize ErrP responses beyond this initial dichotomy, incorporating dimensions such as error awareness, similarity, and severity [53, 60].

Fundamentally, ErrPs are elicited when the brain responds to the occurrence of an error. Building on this principle, researchers have sought to differentiate the types of ErrPs based on the origin and context of the error. One widely recognized distinction is between errors committed by the subject, often referred to as outcome or response errors, and those generated by external systems. For instance, execution errors arise when a machine fails to correctly interpret or implement a user's command, thereby producing a mismatch between intention and system output [54, 61, 62]. Another important category is the feedback error, which occurs when the subject receives feedback that highlights an error, including errors that may have been committed unconsciously during task performance. This type of ErrP is linked to evaluative processes in the brain that update behavior in light of new information [53, 54, 63]. Furthermore, ErrPs can be evoked in observational contexts, whereby an individual detects an error committed by another human or a machine. These observation errors demonstrate that ErrP responses are not limited to self-generated mistakes but extend to error monitoring in interactive or supervisory environments [53, 54, 63].

In addition to these categories, Ferrez and Millán identified a distinct form of ErrP known as the interaction ErrP. This response emerges during direct engagement with an interface when the system generates an erroneous output, thereby disrupting the natural flow of human-machine interaction [37]. The recognition of this category has underscored the importance of considering ErrPs in dynamic, real-time communication scenarios.

The ability to identify and classify ErrPs across these diverse contexts—ranging from self-generated to externally observed and system-mediated errors—has significantly advanced the field. As highlighted by Spüler et al. and others, the successful delineation of these

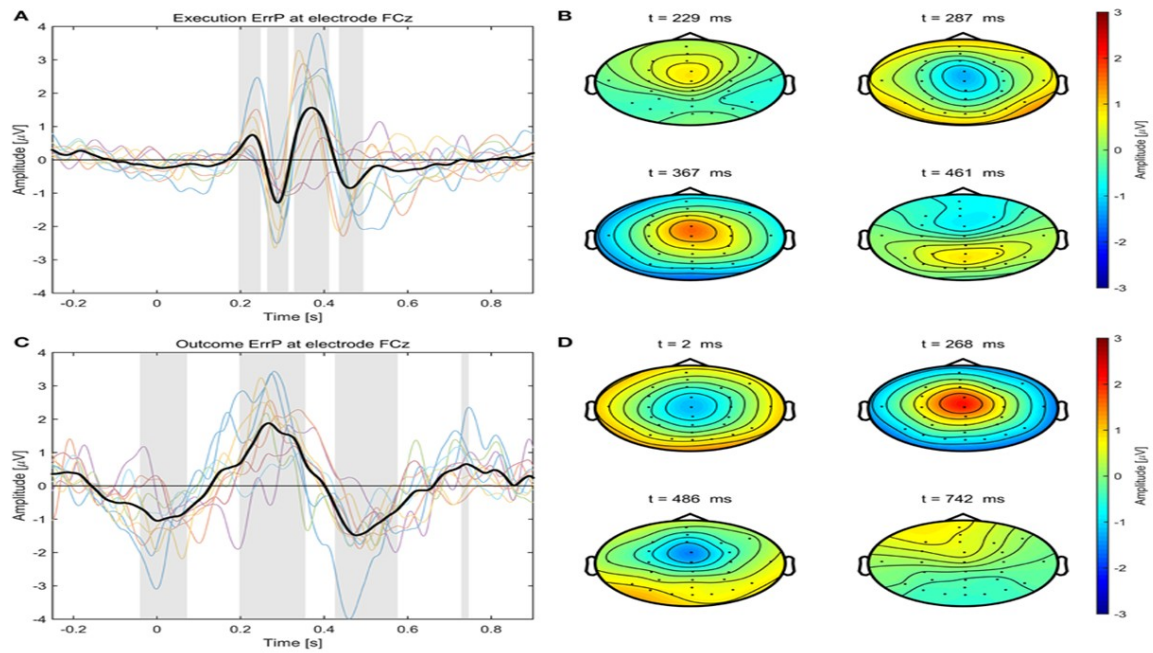


Figure 2.4: Error-related potentials associated with execution and outcome errors. Left: Error-related potentials recorded at electrode FCz for execution errors (A) and outcome errors (C). ErrPs were obtained by computing the difference between error trials and NoError trials. Colored lines represent individual participants, while the bold black line indicates the grand average across all participants. Error occurrence is aligned at $t = 0$ ms. Gray shaded regions indicate time intervals with significant differences between error and NoError trials ($p < 0.05$, Bonferroni corrected). Right: Scalp topographies illustrating the spatial distribution of the error-related potentials for execution errors (B) and outcome errors (D) at the time points corresponding to the maximum deflection within each significant time interval [53]

categories has bolstered the potential for applying ErrP-based mechanisms in brain–computer interface (BCI) systems [53, 62]. This refined understanding strengthens the prospect of designing BCIs that are not only more adaptive but also capable of robust error detection across multiple interaction modalities. Ultimately, the growing taxonomy of ErrPs reflects both the complexity of human error monitoring and its value for optimizing neurotechnological applications.

2.3.5 Commonly Used Classification Algorithms, and Extracted Features of Some Selected ErrP Studies

After conducting a comprehensive review of the existing literature on algorithms applied in electroencephalographic (EEG)-based Error-Related Potential (ErrP) studies, it was observed

that a variety of computational approaches have been employed to enhance the reliability and accuracy of detecting ErrPs. These approaches typically involve three key stages: signal pre-processing, feature extraction, and classification.

The pre-processing stage is critical for improving the quality of the EEG signals by reducing noise and artifacts that may obscure relevant neural activity. Common techniques include band-pass filtering, artifact rejection methods (e.g., Independent Component Analysis for ocular or muscular artifacts), and signal normalization.

Following pre-processing, features are extracted to capture the relevant temporal, spectral, or spatial characteristics of the EEG signals associated with ErrPs. Feature extraction methods reported in the literature include time-domain averaging, spectral power analysis, common spatial patterns (CSP), wavelet transforms, and principal component analysis (PCA). The choice of features plays a decisive role in determining the discriminability of ErrPs from background EEG activity.

In the final stage, machine learning classifiers are employed to categorize the extracted features into error or non-error responses. A range of classifiers have been reported across studies, including Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Gaussian classifiers, k-Nearest Neighbors (kNN), and deep learning-based approaches. The choice of classifier is often influenced by dataset size, computational constraints, and the complexity of the feature space.

Table 2.4 summarizes the signal pre-processing methods, feature extraction techniques, and classifiers employed in selected EEG studies, providing a comparative overview of methodological choices and their application in ErrP detection research. This synthesis highlights both the diversity of approaches and the emerging trends in the field, serving as a basis for the methodological framework adopted in this study.

2.3.6 Relevance of ErrP Studies in BCI Systems Development

The advancement of various research studies towards the classification of ErrP on single and multi-trial basis have opened up a lot of research opportunities for the modification of the present-day BCI systems for improved performances. Some studies have successfully proved that ErrP can be used for error recognition and classifications[36], error awareness and consciousness , improved spelling speed and bit rate [36–38, 52]. Other studies have been developed on the use of ErrP-based BCIs for the rehabilitation of motor-impaired patients using assistive devices such as robots, spellers, gesture recognition systems, and brain-

Table 2.3: Common classification algorithms used in EEG data analysis

Algorithm	Core Idea	Pros	Cons	Typical EEG Use
Linear Discriminant Analysis (LDA)	Finds linear projection maximizing class separability	Fast, simple, works well with few features	Struggles with non-linear boundaries	Motor imagery classification with CSP features
Logistic Regression	Models class probability via sigmoid function	Interpretable, probabilistic outputs	Linear decision boundary only	Binary EEG tasks (e.g., seizure detection)
Support Vector Machine (SVM)	Maximizes margin between classes, with kernel option for non-linearity	Handles high-dimensional data, robust	Can be slow with large datasets	Motor imagery, multi-class mental state classification
Decision Tree	Splits features into homogeneous groups using rules	Easy to interpret, non-linear boundaries	Overfitting risk if deep	Quick prototyping of EEG decision rules
Random Forest	Ensemble of decision trees with random feature selection	Reduces overfitting, handles noisy data	Less interpretable than single tree	Sleep stage classification, workload detection
k-Nearest Neighbors (k-NN)	Classifies based on closest k samples	Simple, no training phase	Slow at prediction, sensitive to scaling	Mental state detection with small datasets
Naïve Bayes	Applies Bayes' theorem assuming feature independence	Very fast, works with small data	Independence assumption often violated	Frequency band power classification
Multilayer Perceptron (MLP)	Feed-forward neural net with hidden layers	Models complex relationships	Needs tuning, overfits small datasets	EEG-based cognitive workload estimation
Convolutional Neural Network (CNN)	Learns spatial/temporal patterns via convolution	Works directly on raw or transformed EEG	Needs large datasets, high computation	EEG spectrogram or topomap classification
Gaussian Mixture Model (GMM)	Models distribution as a sum of Gaussians	Soft classification, probabilistic	Assumes Gaussian distribution	EEG source localization, pattern detection

controlled wheelchairs [64]. Many other researchers have also channeled their interests into using ErrP signals for BCI systems error detection and control, ensuring that human intents and machine commands are not executed if ErrP is detected [37, 38]. This, of course, has significantly improved the overall performance of BCI systems. However, even though some research studies have also found loopholes in this concept [65], its advantages have also been found to have generally outweighed its shortcomings [64].

2.3.7 Methods for Different Applications of ErrP in BCI Real-Time System Control

Given the numerous advantages of error-related potentials (ErrPs) as previously highlighted, several researchers have explored diverse strategies for integrating ErrP signals into brain-computer interface (BCI) systems with the goal of enhancing performance and reliability [66]. A central theme in this body of work is the use of ErrP detection as a corrective mechanism, enabling the system to identify and cancel erroneous actions in real time. Such an approach has been shown to improve the efficiency of information transfer, typically measured in bits per trial (BpT), thereby enhancing the overall throughput of BCI systems [37].

Other researchers have argued for more flexible integration strategies, suggesting that the corrective or adaptive measures employed should depend not only on the type of error detected but also on the specific BCI paradigm being utilized [62]. For example, Combaz et al. proposed an innovative approach in the context of a P300 speller system, whereby the system automatically selects the second-best ranked letter whenever an error is detected. This method demonstrated the potential for ErrP-driven error correction to improve usability and accuracy in communication-based BCIs [67].

Beyond error cancellation and adaptive response strategies, another major avenue of research focuses on enhancing the accuracy and robustness of machine learning classifiers used in BCIs. By leveraging ErrP signals, classifiers can be dynamically tuned to improve performance. For instance, Jrad et al. demonstrated that in support vector machine (SVM)-based systems, the regularization parameter—responsible for balancing the trade-off between model complexity and empirical error—can be adaptively adjusted following error detection. This adaptive tuning resulted in measurable gains in system performance [39, 62, 68–71].

Taken together, these findings highlight the versatility of ErrP integration strategies, which range from direct error cancellation to adaptive system responses and classifier optimization.

Such approaches not only improve accuracy and speed but also enhance the overall user experience, positioning ErrP-based error monitoring as a critical component in the development of next-generation BCI systems [62, 72].

2.4 Chapter Summary

This chapter provides a comprehensive review of existing literature and studies relevant to the present research. The discussion begins with an exploration of human-machine interaction, focusing particularly on the concept of trust between humans and machines, as well as the reciprocal influence of human behavioural dynamics on machine performance and adaptation during interactive processes.

The review further examines the role of Brain-Computer Interface (BCI) technologies, with emphasis on the mechanisms that underpin brain signal generation and external stimulation. In this context, particular attention is given to methodologies for brain signal extraction, processing, and interfacing, which are essential for isolating and analyzing neural activities of interest.

Additionally, the chapter highlights the fundamental components of electroencephalography (EEG), outlining the major approaches to EEG signal classification. Special focus is placed on event-related potentials, including the Error-related Potential (ErrP) and the P300 response, both of which play a critical role in enhancing the accuracy and reliability of BCI systems. Furthermore, various classification algorithms employed in EEG analysis are reviewed, illustrating their strengths, limitations, and applicability to different experimental contexts.

Through this synthesis, the chapter establishes a foundational understanding of the theoretical and technological frameworks that support the subsequent experimental investigations in this study.

Table 2.4: Extracted features and classification algorithms for some Errp studies

Extracted Features	Classification Algorithms Used	Title of Article
Spatio-temporal	<i>Shrinkage LDA</i>	<i>Online Detection of ErrP boosts of the performance mental typewriter</i>
Time, Frequency and both	<i>SVM</i>	<i>ErrP during Continuous Feedback, Using EEG to detect errors of different types and severity</i>
Time	<i>(MLP ANN),and (FCM) classifiers</i>	<i>Classification of ErrP Associated Response Error in Actors and Observers Based Autoregressive Modelling</i>
Time and Frequency	<i>SVM with 2nd order polynomial Kernel</i>	<i>The Effect of Vibrotactile Feedback on ErrP based Adaptive Classification of Motor Imagery</i>
Time	<i>Gaussian Classifier</i>	<i>Learning from ErrP in Non-Invasive Brain-Computer Interfaces</i>
Time	<i>(SWLA)</i>	<i>"You Have Reached Your Destination: A Single EEG Classification Studies"</i>
Time and Frequency	<i>LibSVM (SVM+ RBF Kernel)</i>	<i>Classifying response correctness across Different Task: A Machine Learning Approach</i>

Analysing and Modeling of Human Trust to a Navigation Robot

3.1 Chapter Introduction

Human trust in machines has been conceptualized in different ways across disciplines. For instance, the meaning of trust differs significantly in interpersonal, organizational, and human-machine contexts [3, 10, 14]. Trust has also been categorized into dispositional, situational, and learned components [14]. Dispositional trust is shaped by relatively stable demographic factors such as gender or cultural background, whereas situational and learned trust vary dynamically with factors such as task complexity, user experience, and self-confidence. Beyond these definitions, many authors describe human trust in machines as a key factor underpinning effective collaboration between humans and their machine counterparts [10–12, 17, 43, 73–75].

Several studies have investigated how human trust develops and fluctuates in human-robot interaction. For example, Aiken et al. showed that understanding a robot's limitations, rather than its task intentions, fosters user confidence and reliability in collaborative tasks [16, 17]. However, the nature and severity of a robot's errors strongly influence human trust. Salem et al. demonstrated that task-related errors can diminish user confidence [19], while other work

has shown that a robot's ability to acknowledge mistakes and attempt recovery can restore trust [64]. Furthermore, human perceptions of error severity vary: some individuals may perceive a given mistake as critical, whereas others regard it as minor, depending on context [76]. Additionally, Jung et al. investigated trust in agents with either human-like faces (HF) or robotic faces (RF) during decision-making tasks. Participants consistently reported higher trust in HF agents, regardless of their risk-taking personality traits [2].

The role of system reliability has also been investigated in automated decision-making. De Vries et al. studied error effects in a computer-based route-planning system. Their findings revealed that high automation error rates (AERs) decreased system trust, while high manual error rates (MERs) reduced user self-confidence, highlighting the dual influence of machine and human performance on trust formation [77].

Machine-induced errors are unavoidable and play a central role in shaping trust in human-machine interactions (HMLs). Both humans and machines are prone to mistakes, which may vary in severity and relevance to the task. Severe errors often draw user attention and demand corrective intervention, while minor errors may be overlooked. Effective human-machine collaboration therefore requires a foundation of mutual trust and understanding. Importantly, machine errors are typically associated with algorithmic limitations, task complexity, or system design flaws. Thus, understanding error types, causes, and user perceptions is essential when investigating trust in HMLs.

Trust is also context-dependent, influenced by organizational needs and specific application domains [10–12, 20, 78–83]. Different industries require machines to perform distinct tasks, making it difficult to establish a generalized trust model across disciplines [50, 84, 85]. While universal frameworks have been proposed, trust formation remains highly dependent on task demands and organizational priorities [74, 75]. Misaligned expectations about machine performance can lead to misuse or disuse of systems, further emphasizing the importance of contextualized trust models. Despite increasing research interest, trust in HMLs under real-world industrial conditions remains underexplored [3, 86–89].

Building on these insights, this chapter investigates factors influencing human trust in a machine using a simulated robotic arm designed to mimic a factory machine navigating through multiple objects to identify a target. Three experimental scenarios were developed, each involving different numbers of objects, error rates, and error types. Additionally, participants were given opportunities to intervene and correct errors made by the robot.

This chapter addresses three primary research questions: (i) how variations in error rates influence human trust in machines, (ii) how changes in system speed and accuracy affect

trust, and (iii) whether the opportunity to intervene and correct severe errors alters users' trust in machine performance. Finally, a regression model was proposed to estimate and predict human trust under these conditions.

Findings of this chapter have been published in the 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering, MetroXRINE 2023 - Proceedings [90]

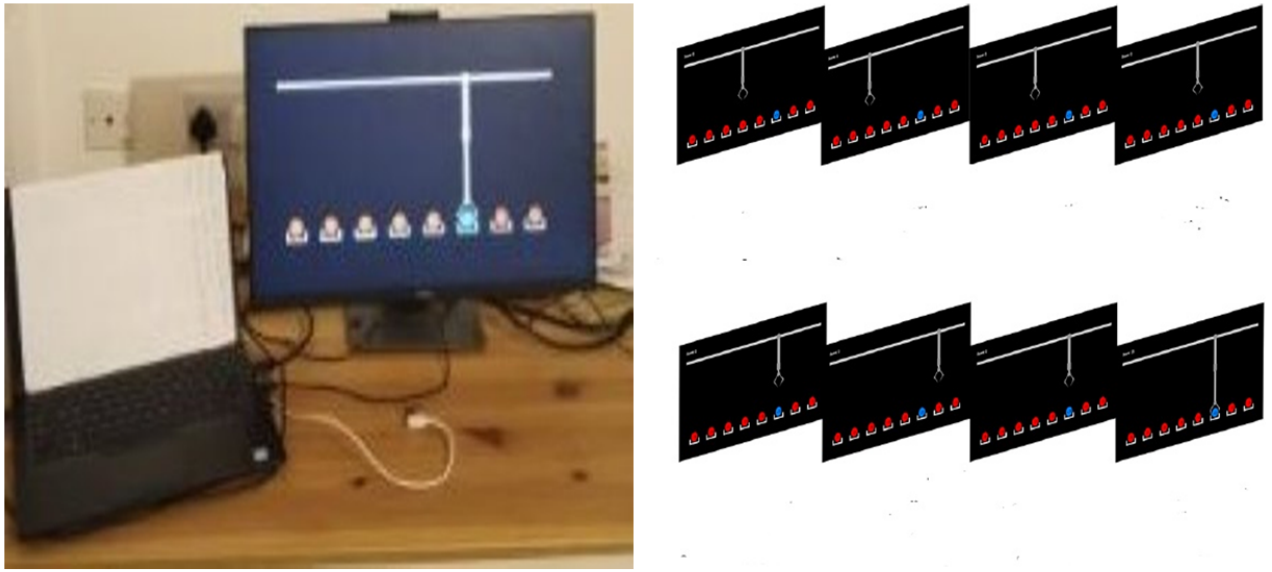


Figure 3.1: Screen shot of the Simulated robotic claw performing some of the task

3.2 Methods

3.2.1 Participants

Ten adult participants (mean age = 27 ± 9 years; 5 male and 5 female) were recruited for the experiment. Prior to data collection, ethical approval was obtained from the Ethics Committee of the Department of Automatic Control and Systems Engineering at the University of Sheffield. The study was conducted at the University of Sheffield, UK. Upon arrival at the laboratory, each participant was seated individually in front of a computer monitor and instructed to observe the movements of a simulated robotic arm performing a factory-related

task. Participants were asked to follow the on-screen instructions carefully while monitoring the robotic arm's actions throughout the session.

Notably, the limited sample size in this study was primarily a consequence of the COVID pandemic, which disrupted normal research activities and significantly constrained participant recruitment. Public health restrictions and safety considerations reduced the number of individuals who could be invited and safely involved in the experimental sessions.

3.2.2 Experiment

To address the study aims, an experiment was designed to simulate the behavior of a robotic arm performing a factory task, where the arm navigated through objects to identify and pick up a designated target. This setup was chosen because it allowed the introduction and control of a finite number of potential errors.

The robot could perform six possible movements: three correct and three incorrect. Correct movements were defined as: Move Towards Target (MTT): the robot moved closer to the designated target. Target Reached (TR): the robot arrived at the target location. Correct Target Identified (CTI): the robot successfully identified the designated target.

Conversely, incorrect movements were classified as errors, consisting of: Move Away (MA): the robot moved further away from the designated target. Stepped Off Target (SOT): the robot reached the correct target but failed to identify it, and Wrong Target Identified (WTI): the robot incorrectly identified or grasped one of the distractor objects (red balls). See Figure 3.1

This classification framework enabled systematic observation of both successful actions and errors, providing a structured means to assess their impact on participants' trust in the robot.

The speed-accuracy trade-off was manipulated directly by operating the system in either a fast or slow mode, while accuracy was manipulated independently through adjustments to the probability of specific error types. The robot operated at a lower speed with higher accuracy when fewer severe errors (WTI errors) were generated, and at a higher speed when the frequency of WTI errors increased. This dual manipulation enabled a clear investigation of how variations in system speed and accuracy (error likelihood) influenced participants' trust in the robot.

The scenarios were not presented in a fixed order. At the beginning of each experiment,

the experimental blocks were randomized to minimize potential order effects. For example, the sequence of blocks could vary such that a participant completed Block C before Blocks A or B. Additionally, different blocks of conditions were interleaved to further reduce systematic sequencing influences.

Trust was assessed through discrete ratings provided by participants during their continuous observation of the robot. At predetermined intervals, the system prompted participants to enter a trust rating on a 1–5 scale, where higher values indicated greater trust (High Trust ratings = 4; Low Trust ratings = 2; mid-range rating = 3, excluded from analysis). Participants recorded their responses using a designated key on the keyboard. This procedure helped maintain participants’ attention and engagement throughout the experiment.

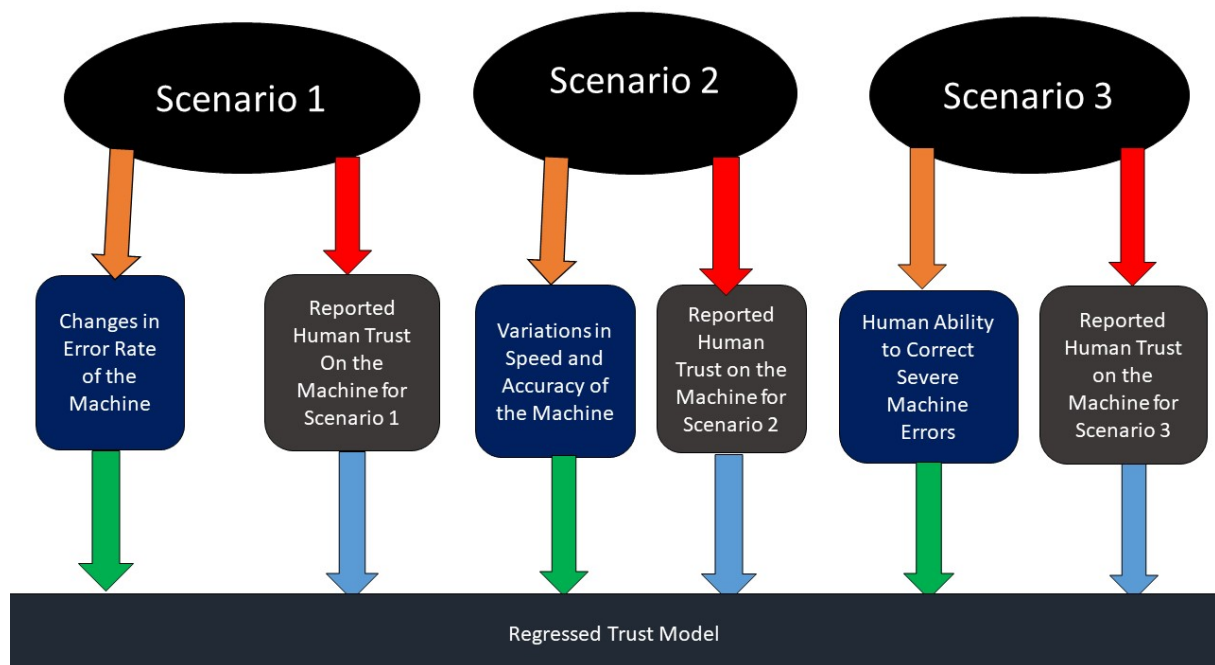


Figure 3.2: Experimental Block Flow Diagram

3.2.3 Task

At the start of each task, the program automatically designated a blue ball as the target destination. The robotic claw was initialized at a position no more than three red balls away from the target (Figure 3.1). The robot’s objective was to navigate toward the blue ball, reach the target, and grasp it.

Errors were categorized into three types:

- Move Away (MA): the robot moved away from the target.
- Stepped off Target (SOT): the robot reached the target but stepped off it.
- Wrong Target Identified (WTI): the robot incorrectly identified or grasped one of the red balls instead of the blue ball.

A task was considered complete when the claw grasped either the blue ball or one of the red balls, after which it automatically reset to begin a new run. This process was repeated until the predetermined number of tasks had been completed. Across the experiment, the claw executed approximately 120 tasks.

The program was configured to pause after every five tasks to allow participants to report their trust in the robot's performance. Trust ratings were provided on a five-point scale, entered using the computer keyboard, where 1 indicated low trust, 3 indicated moderate trust, and 5 indicated high trust. For example, a rating of "1" reflected a low level of trust in the observed actions of the robot, while a rating of "5" reflected high trust.

In Scenario Three, when the claw committed a WTI error, the program automatically paused, and a message was displayed instructing the participant to press a key to resume the task. This ensured that participants directly acknowledged the error before continuing.

The experiment was conducted in a quiet, controlled environment. Participants were asked to turn off their mobile phones to minimize distractions. The experimental design comprised seven blocks: three blocks varying error probability (Low, Medium, and High), two blocks varying error severity (Low and High), and two blocks examining human ability to intervene, in which the error probabilities were kept constant across blocks. Each participant observed at least six of the seven experimental blocks, resulting in a minimum of 240 trials per participant. The specific blocks presented to each participant were randomised. Consequently, a participant might, for example, first observe the penultimate block, followed by a selection of intermediate blocks or the final block.

To avoid bias, participants were not informed of the study design or hypotheses. Instead, they were instructed to carefully observe the robot's movements and count the number of errors committed during each run.

At the conclusion of the experiment, data on both the robot's task performance and participants' trust ratings were collected for subsequent analysis in line with the study hypotheses.

3.2.4 Error Definition

Errors such as Move Away from Target (MA), Step Off Target (SOT), and Wrong Target Identified (WTI) differ in severity and consequence, making them useful for examining how humans form or adjust trust in robots. The error severity ranking was determined based on the potential consequences of the error on task performance and the level of recovery required. Low-severity errors represent minor deviations that do not disrupt task completion and can be automatically corrected by the robot. Medium-severity errors introduce inefficiencies or temporary disruptions during task execution but still allow successful task completion. High-severity errors result in significant task failure or require human intervention for recovery. This classification aligns with established risk assessment principles, where error severity is defined by the impact of the failure rather than its probability of occurrence.

MA errors represent operational inefficiencies in the robot's task execution rather than critical failures. These errors occur when the robot deviates from its intended or optimal trajectory while performing a task. For example, the robot may take a slightly incorrect path, hesitate, or execute an imprecise movement before correcting itself. Although such deviations are noticeable to an observer, they do not typically prevent the task from being completed successfully. In many cases, the robot is able to self-correct or the issue can be easily resolved with minimal intervention. Because MA errors do not lead to severe consequences or task failure, their impact is generally limited to a reduction in perceived efficiency or competence. However, repeated occurrences may still influence a human observer's perception of the robot's reliability, leading to a moderate reduction in trust, as the system appears less precise or predictable than expected.

SOT errors are relatively minor in severity and occur when the robot exhibits slight deviations during task execution. These deviations do not typically compromise task completion but may be noticeable to users as small imperfections in the robot's performance. As a result, users are likely to be more tolerant of such errors, and their occurrence is expected to have only a limited impact on overall trust in the system. An example of an SOT error can be observed in a robotic pick-and-place task in a warehouse or assembly setting. For instance, when the robot attempts to grasp an object from a conveyor belt, it may momentarily misalign its gripper and make a slight adjustment before successfully picking up the item. The robot still completes the task, but the movement appears slightly imprecise or less smooth than expected. Another example could occur during a navigation task. A mobile robot moving along a predefined path might drift slightly from the centre of the path and then automatically correct its trajectory after a brief adjustment. Although the deviation is visible to an observer, the robot quickly returns to the intended path without affecting task

completion. In both cases, the robot's behaviour demonstrates a minor deviation rather than a critical failure, and observers are likely to interpret the behaviour as a small imperfection in performance rather than a serious problem. Consequently, such errors are unlikely to significantly reduce users' overall trust in the system.

In contrast, WTI errors are critical because they directly compromise task success by misidentifying the intended target, often requiring human intervention. These high-severity errors are more likely to erode trust significantly.

By comparing user trust across these distinct error types, researchers can capture how people weigh the importance of error severity and task relevance in human-machine interactions. This differentiation provides insight into which types of mistakes are most detrimental to trust and which may be tolerated within collaborative contexts.

Following completion of the experiment, I confirmed that the implemented error rates were consistent with the values predefined for each block. To achieve this, the actual occurrences of the three error types (MA, SOT, and WTI) were recorded across all scenarios and blocks. The observed error rates were then calculated and compared against the experimental design specifications to ensure alignment with the intended conditions (see equations below)

$$MAerrorrate = \frac{(MA)}{(MA + MTT + WTI)} \quad (3.1)$$

$$SOTerrorrate = \frac{SOT}{(SOT + CTI)} \quad (3.2)$$

$$WTIerrorrate = \frac{WTI}{(WTI + MA + MTT)} \quad (3.3)$$

3.2.5 Data Analysis

To align the analysis with the study hypotheses, participants' reported trust ratings were recorded alongside the calculated error rates for each error type across all experimental blocks.

First, paired t-tests were conducted to examine differences in error rates across blocks within each scenario, ensuring consistency with the experimental design specifications.

In Scenario One, the relationship between trust and MA errors was assessed using Pearson correlation analyses, with average reported trust values correlated against the frequency of MA errors.

In Scenario Two, average trust ratings were compared between the two blocks, which differed in their emphasis on speed versus accuracy (Table 3.2). Paired t-tests were conducted to determine whether the probability of WTI errors—and thus the speed-accuracy trade-off—produced significant differences in trust.

In Scenario Three, participants reported their trust levels each time they intervened to correct WTI errors. Pearson correlation analyses were then performed to evaluate the relationship between average trust ratings and WTI error frequencies across blocks.

Finally, to develop a general model of trust across conditions, trust ratings from all three scenarios were pooled and regressed against the total number of errors committed in each case. This allowed estimation of trust as a function of error type and frequency across the full experimental design.

Scenario One

The objective of Scenario One was to examine how variations in error rates influence human trust in the robot. To achieve this, the experiment was structured into three blocks of robot actions, each lasting three minutes. Within each block, predefined probabilities were assigned to the three types of errors (Table 3.1).

In Block 1, the error probabilities were set at 0.10 for MA, 0.20 for SOT, and 0.10 for WTI. In Block 2, the probability of MA errors was increased to 0.25, while the probabilities for SOT (0.20) and WTI (0.10) remained constant. In Block 3, the MA error probability was further increased to 0.40, with SOT and WTI probabilities unchanged. Thus, only the probability of MA errors was manipulated across the three blocks, while SOT and WTI error rates were held constant.

The decision to vary MA errors exclusively was based on their relative importance in the navigation task. An MA error occurs when the robot moves away from the designated target, which greatly reduces the likelihood of successfully completing the task. By contrast, SOT and WTI errors were held constant to avoid confounding effects, ensuring that observed changes in trust could be attributed primarily to MA error variations.

To validate this design, Pearson correlation analyses were conducted to assess the relationship between error rates and trust ratings. Results revealed a significant negative correlation between experienced MA errors and trust, whereas no significant correlations were observed for SOT or WTI errors across the blocks.

Table 3.1: Changing MA error rate across blocks in Scenario One. In this table, SOT and WTI error rates remain constant across blocks while MA progressively increases

"Note: MA= Move away error; SOT = stepped off Target error and WTI Wrong Target Identified error"

Blocks	MA Errors	SOT Errors	WTI Errors
Block 1	0.10	0.20	0.10
Block 2	0.25	0.20	0.10
Block 3	0.40	0.20	0.10

Scenario Two

The aim of Scenario Two was to investigate how human trust in the robot changes when the system's operational priority shifts between speed and accuracy in identifying the target. For this purpose, the experimental design emphasized WTI errors as the most critical error type, since the task was considered complete whenever the robot grasped an incorrect object.

The experiment consisted of two blocks, each lasting three minutes, with error probabilities defined as shown in Table 3.2. In Block S/A1, the error probabilities were MA = 0.30, SOT = 0.30, and WTI = 0.10. In Block S/A2, the probabilities were MA = 0.10, SOT = 0.10, and WTI = 0.50.

The design of these blocks enabled systematic manipulation of the trade-off between speed and accuracy. In Block S/A1, the robot committed fewer WTI errors, reflecting lower speed but higher accuracy. Conversely, in Block S/A2, the increased WTI error probability reflected higher speed but lower accuracy. Thus, WTI errors served as the key parameter to model performance variation between "fast but inaccurate" and "slow but accurate" robotic behavior.

This manipulation parallels industrial scenarios in which machines may be configured to prioritize either speed or accuracy in task execution. The focus of this scenario was therefore to assess how these differences influence participants' trust in the robot's performance.

Scenario Three

The objective of Scenario Three was to examine the effect of human interaction and control on trust in the robot. To address this, the experiment was divided into two blocks, with error

Table 3.2: preset Error Rates for speed and robot accuracy in scenario Two

No. of Blocks	<i>MA Errors</i>	<i>SOT Errors</i>	<i>WTI Errors</i>
BlockS/A 1	<i>0.30</i>	<i>0.30</i>	<i>0.10</i>
BlockS/A 2	<i>0.10</i>	<i>0.10</i>	<i>0.50</i>

Table 3.3: Error probabilities of the three types of Errors across the blocks in scenario Three

Blocks	<i>MA Errors</i>	<i>SOT Errors</i>	<i>WTI Errors</i>
Block 1	<i>0.40</i>	<i>0.20</i>	<i>0.10</i>
Block 2	<i>0.40</i>	<i>0.20</i>	<i>0.10</i>

probabilities fixed at MA = 0.40, SOT = 0.20, and WTI = 0.10 for both blocks (Table 3.3). Thus, the robot committed the same number of errors across the two conditions.

Unlike Scenarios One and Two, which focused on error rate manipulation, Scenario Three introduced a shift in the experimental paradigm by allowing participants to actively intervene and correct errors committed by the robot. In particular, each time the most severe error—the Wrong Target Identified (WTI) error—was committed, participants were prompted on the computer to intervene and correct it. During these trials, participants were instructed to report their level of trust in response to the robot’s observed actions while also exercising corrective input when necessary.

This design enabled investigation of trust development in a collaborative context, where human participants acted as team players alongside the robot to achieve the task goal.

Table 3.4: Pearson’s correlation results between reported trust and numbers of errors across Blocks in Scenario One

	<i>MA Errors</i>	<i>SOT Errors</i>	<i>WTI Errors</i>
Pearson Correlation	<i>-.493</i>	<i>-.057</i>	<i>.001</i>
Sig. (2-tailed)	<i>.000</i>	<i>.585</i>	<i>.992</i>
N	<i>94</i>	<i>94</i>	<i>94</i>

3.3 Results

3.3.1 Scenario One

To examine the influence of error types on participant responses, a paired-samples t-test was conducted to compare the mean differences in error frequencies across the experimental blocks. The analysis considered three distinct error categories: Move Away (MA), Step Off Target (SOT), and Wrong Target Identified (WTI).

For each error category, the mean frequency of errors within each block was first calculated across all participants. These mean values were then compared across blocks using paired comparisons. Specifically, the mean frequency of MA errors in Block 1 was paired with the corresponding mean frequencies in Blocks 2 and 3. Similarly, the mean frequencies of SOT errors in Block 1 were paired with those in Blocks 2 and 3, and the same procedure was applied to the WTI error category. This process was repeated such that equivalent error types across all blocks were systematically paired and compared.

This approach enabled the assessment of whether the frequency of each specific error type differed significantly between experimental blocks while accounting for within-subject variability.

The results indicated a statistically significant difference in the mean values of MA errors across experimental blocks ($p < .05$), suggesting that the frequency of MA errors varied significantly throughout the course of the experiment. This observation is consistent with the predefined error probabilities specified during the experimental design stage. In contrast, the mean values of SOT and WTI errors did not show statistically significant differences across blocks ($p > .05$), indicating that these error types remained relatively stable across the experimental conditions. These findings suggest that MA errors were more sensitive to block-specific manipulations or task conditions than SOT and WTI errors.

Furthermore, the analysis of variance (ANOVA) results corroborated these findings. A statistically significant difference was observed in MA errors across blocks ($p = .04$), whereas no significant differences were found for SOT errors ($p = .86$) or WTI errors ($p = .39$). Collectively, these results indicate that only MA errors demonstrated systematic variation across blocks, while SOT and WTI errors remained largely unaffected by the block structure of the experiment.

Building upon these findings, a Pearson correlation analysis was conducted to explore the

relationship between the frequency of MA errors and participants' reported trust in the robot. For this analysis, MA error counts were aggregated across all three blocks for each participant, and these values were correlated with corresponding trust ratings. The analysis revealed a significant negative correlation ($r = -0.493$, $p < .001$), indicating that higher frequencies of MA errors were associated with lower levels of trust in the robot. No significant result was found for SOT and WTI errors as shown in 3.4.

This result provides empirical support for the hypothesis that MA errors—where the robot incorrectly attributes actions or intentions—have a detrimental effect on user trust. The negative correlation suggests that participants perceived MA errors as more disruptive or misleading, thereby undermining their confidence in the robot's reliability and decision-making capabilities. These findings underscore the importance of minimizing MA errors in human-robot interaction systems to preserve user trust and engagement.

Table 3.5: Pearson's Correlation Results of Scenario Two

	<i>MA Errors</i>	<i>SOT Errors</i>	<i>WTI Errors</i>
Pearson Correlation	.484	-.237	-.582
Sig. (2-tailed)	.111	.459	.047
N	16	16	16

3.3.2 Scenario Two

To investigate the effect of speed and accuracy on human trust in the robot, a paired samples t-test was again conducted to compare reported trust between the low-speed/high-accuracy condition and the high-speed/low-accuracy condition. Results indicated that participants reported significantly higher trust in the robot when it operated at lower speed with higher accuracy compared to when it operated at higher speed with lower accuracy. Specifically, the mean trust score was higher in the low-speed/high-accuracy block ($M = 3.38$, $SE = 0.263$) than in the high-speed/low-accuracy block ($M = 2.50$, $SE = 0.327$), $t(7) = 2.966$, $p < .05$. This finding suggests that accuracy was a stronger determinant of trust than speed, as trust decreased when the robot performed tasks more quickly but with greater errors.

To further examine the role of error types in shaping reported trust, Pearson's correlation analyses were performed. Reported trust was found to be significantly negatively correlated

with WTI errors across the two blocks, $r = -.582$, $p < .05$. This indicates that increases in WTI errors were associated with reductions in trust.

Additional correlation analyses were conducted to test the relationship between trust and other error types (MA and SOT errors). No significant associations were observed in either case ($p > .05$) as shown in Table 3.5. This result suggests that although the robot committed multiple error types, only WTI errors significantly contributed to the observed changes in human trust.

Table 3.6: Pearson's Correlation Results of Scenario Three

	<i>MA Errors</i>	<i>SOT Errors</i>	<i>WTI Errors</i>
Pearson Correlation	.133	.451	-.434
Sig. (2-tailed)	.599	.060	.072
N	18	18	18

3.3.3 Scenario Three

To examine the influence of human intervention in correcting the robot's most critical errors on reported trust, a Pearson correlation analysis was conducted between trust scores and WTI errors. This analysis included data from six participants, as four participants did not take part in this phase of the experiment. WTI errors were selected for analysis because they trigger false alarms whenever the robot commits such errors.

The results indicated no significant correlation between reported trust and WTI errors ($p > .05$). This finding suggests that WTI errors did not have a measurable impact on participants' trust in the robot during this scenario. One possible interpretation is that participants may have been able to maintain or regain their trust in the robot through successful collaboration, where they could effectively intervene to correct errors and work jointly with the robot to achieve the task goal.

3.3.4 Estimating Human Trust Using Regression Analysis

In general, this result further demonstrates that understanding human levels of trust in a machine in complex situations of uncertainties and complexities where different mistakes of the machine could occur would aid humans or users of the machine to further make informed

decisions that will enable team partnerships to manipulate the systems to achieve a setup goal. This is important to avoid complete failure in the event of uncertainties.

Building on the correlation analyses conducted across scenarios, a multiple regression analysis was performed to examine whether different error types could be used to predict reported trust. Error rates from Scenarios One and Two were pooled and regressed against participants' reported trust scores.

The results indicated that both MA and WTI errors were significant predictors of trust, $F(3, 43) = 8.084$, $p < .001$, with regression coefficients of $B = -5.515$ (MA errors) and $B = -3.507$ (WTI errors), $t = 10.474$, $p < .001$. The model explained approximately 36.1% of the variance in reported trust ($R^2 = .361$). In contrast, SOT errors were not significantly associated with trust ($p > .05$), consistent with the earlier correlation analyses that showed no meaningful relationship between SOT errors and trust.

Overall, these results suggest that trust in the robot is strongly influenced by MA and WTI errors, but not by SOT errors. This finding highlights the importance of error type in shaping human-machine trust dynamics. More broadly, the ability to model and estimate trust in contexts characterized by uncertainty and multiple error sources may support the development of adaptive human-robot partnerships, enabling users to make informed decisions and sustain effective collaboration even under challenging conditions.

3.3.5 Discussion

The results of this study highlight that human perceptions of machine behaviour are not uniform but vary considerably among individuals. This variation aligns with previous findings suggesting that trust in automation is influenced by subjective factors such as personality traits, prior experience with technology, and cognitive biases [10, 14]. It appears that individual differences play a significant role in how participants interpret robotic actions and assign trust to automated systems.

My finding that MA and WTI errors significantly predict reported trust, while SOT errors do not, aligns with prior research showing that error severity, consequence, and visibility play key roles in human trust calibration [54]. In particular, errors that directly impact task correctness or that violate competence expectations tend to erode trust more strongly than errors of timing or minor inefficiency (Understanding the Effects of Robot Errors in HRI; "Sorry, it was my fault"). Moreover, trust-prediction frameworks (e.g., the bi-directional trust model) posit that systems' capability violations (rather than minor deviations) carry more weight

in trust modeling [19, 91]. These lines of evidence support the logic that human trust is more sensitive to MA and WTI errors, which compromise task outcomes, than to SOT errors, which primarily reflect suboptimal performance without outright failure. Prior studies have similarly indicated that humans tend to tolerate a consistent error rate but react more strongly to unexpected or fluctuating failures, which may be perceived as unreliability or loss of control[18]

Another important observation concerns the trade-off between speed and accuracy. Participants expressed higher levels of trust in a robot that operated at a slower speed but with greater accuracy compared to one that performed faster with more errors. This finding supports the argument that perceived reliability and competence are stronger determinants of trust than efficiency or task completion speed [92, 93]. It also underscores the importance of designing robotic systems that prioritise consistent and accurate performance to maintain user trust, even at the expense of operational speed.

Interestingly, when participants were allowed to intervene and correct severe robot errors, the results revealed no significant correlation between mean trust values and error rates. This suggests that opportunities for human intervention may mitigate the negative effects of robot errors on trust. In other words, the ability to exert control or correct system behaviour appears to restore user confidence in the robot's competence. This aligns with the concept of calibrated trust, where maintaining an appropriate balance between human oversight and automation autonomy supports sustained trust over time [14, 94]. The finding also resonates with theories of shared control in human-robot interaction, which emphasise that empowering users to recover from automation failures can prevent trust erosion and foster resilience in human-robot partnerships [95].

As earlier stated, my regression results further indicated that both MA and WTI errors were significant predictors of trust, $F(3, 43) = 8.084, p < 0.001$, with regression coefficients of $B = -5.515$ (MA errors) and $B = -3.507$ (WTI errors), $t = 10.474, p < .001$. The model explained approximately 36.1% of the variance in reported trust ($R^2 = .361$). In contrast, SOT errors were not significantly associated with trust ($p > .05$), consistent with the earlier correlation analyses that showed no meaningful relationship between SOT errors and trust.

Overall, these findings contribute to a growing body of evidence suggesting that human trust in robotic systems is dynamic, context-dependent, and strongly shaped by perceived reliability and the capacity for human oversight. Future work could further investigate how different modes of human intervention or adaptive feedback mechanisms influence trust calibration, particularly in complex or high-stakes environments.

3.4 Chapter Summary

This chapter addresses the first aim and objective of the project: to investigate how variations in machine error rates, the severity of those errors, and the extent of human ability to intervene in correcting them influence human trust in machines. The work presented here has been previously published in the Proceedings of the 2023, IEEE, MeTROXRaine, Milano, Italy, reflecting its contribution to ongoing scholarly discourse.

The chapter begins by situating this objective within its theoretical and empirical foundations, drawing on prior research in human-machine interaction, trust dynamics, and error management. It reviews a range of experimental studies and methodological approaches that have explored the interplay between system reliability, error characteristics, and user trust, with particular emphasis on the strengths and limitations of these approaches. Special attention is given to identifying how different dimensions of error—frequency, severity, and recoverability—shape user perceptions and trust-related outcomes.

Through a critical synthesis of the literature, this chapter not only highlights the advancements achieved in understanding trust calibration but also identifies persistent gaps and open challenges that motivate further investigation. It underscores the need for a more integrated perspective that combines error-related factors with user intervention mechanisms, thereby offering a clearer framework for assessing trust in real-world applications. By doing so, the chapter provides both the conceptual grounding and empirical justification for the subsequent phases of this project, while establishing the relevance of the present contribution within the broader trajectory of human-machine trust research.

Investigating Human Trust in Machine Using ErrP Signals

4.1 Chapter Introduction

Error-related potentials (ErrPs) are neural signals generated when individuals recognize the occurrence of an error, whether committed by themselves or by external agents such as machines [55]. Previous research has demonstrated that ErrPs can be reliably detected from electroencephalographic (EEG) signals on a single-trial basis, allowing the differentiation between correct and erroneous actions [41, 42, 46, 47, 54]. Wirth et al. further expanded this approach by classifying distinct types of errors against one another [36]. These studies have highlighted the feasibility of employing single-trial error versus non-error [63] or error versus error classifications [36] as feedback mechanisms in reinforcement learning-based brain-computer interface (BCI) systems [67]. Such systems pave the way toward autonomous BCI applications, where machines can adaptively learn appropriate actions based on human assessments of correctness. This approach has shown potential in tasks such as simple navigation [41, 42], reducing cognitive workload and enhancing human-machine collaboration.

Recent advancements also suggest that ErrPs, alongside other EEG markers such as the

P300, could provide insight into variations in human trust toward machines. By presenting comparable information to both humans and machines, and measuring corresponding EEG responses while participants report their trust levels, it becomes possible to model trust dynamically. For example, an algorithm capable of estimating trust in real time (e.g., per minute during task execution) using EEG signals could continuously and accurately track neural responses associated with trust during human-machine interaction.

Although neural correlates of trust have been extensively investigated in human-human contexts [5], it is critical to recognize that trust in human-machine interaction may involve distinct mechanisms under different task conditions. Self-report measures have frequently been used to estimate trust (e.g., [1]), but these methods suffer from limitations, including interruptions to task performance, variability in self-assessment accuracy, participant bias, and cognitive burden [10, 12]. Physiological indicators such as galvanic skin response (GSR) and heart rate variability have also been explored; however, their relatively slow temporal dynamics limit their utility for real-time trust measurement. More recently, EEG studies have revealed links between human confidence and error-related positivity (oPe), though findings remain constrained by experimental design and environmental factors [10].

To address these challenges, EEG presents a promising avenue for assessing trust in machines, owing to its high temporal resolution (on the order of milliseconds), which surpasses that of slower physiological signals such as GSR. Nevertheless, knowledge of the neurocognitive basis of trust in human-machine interaction remains limited, and further research is needed to account for the conceptual complexity of trust [78, 96]. Specifically, relationships between trust and EEG components—including the error-related negativity (ERN) and error positivity (Pe) of the ErrP, as well as the P300 associated with correct actions—require deeper investigation. Additionally, contextual factors such as individual personality traits, task type, user familiarity with the machine, and the operational environment may all influence trust [97–99]. Understanding these relationships is essential to foster reliable and harmonious human-machine collaboration.

In this chapter, these findings are extended by investigating human trust in relation to errors committed by machines, with particular focus on EEG components of the ErrP (ERN and Pe). The study aims to determine whether trust can be distinguished at different levels by analyzing these neural markers. Human perception of machine actions may vary across individuals and domains; however, shared patterns in trust responses could enable robust trust modeling, reducing risks of system misuse or disuse [74].

To explore this, EEG signals were analyzed using a linear discriminant analysis (LDA) classi-

fier, with a support vector machine (SVM) applied for validation. Discriminative time-domain features were extracted from three of the eight electrode sites used in the experiment. Data were collected from Ten adult participants who observed the actions of a robotic claw system displayed on a computer screen (as described in Chapter 3) but seven participants were used for analysis. Participants were not informed of the experiment's purpose but were instructed to observe the system's actions and provide trust ratings based on their perception. The resulting data enabled investigation of the relationship between trust and ErrP components, as well as classification of trust levels (high vs. low).

Findings of this chapter and Chapter 5 have also been published in the 2025 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering, MetroXRAINE 2025 - Proceedings

4.2 Methods

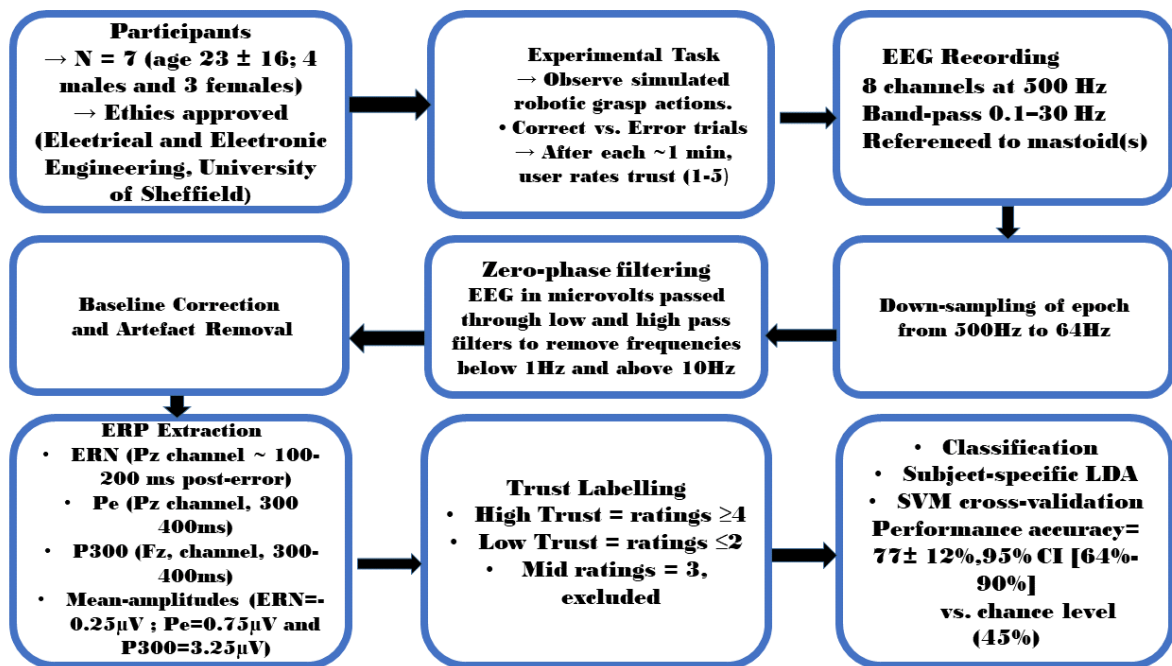


Figure 4.1: Complete Flow Block Diagram of the Experiment

4.2.1 Participants

Initially, Ten healthy adult participants were recruited for this study. However, seven(7) participants (Mean age= 23± 16; 4 males and 3 females) were included in neurophysiological

analysis while three were excluded because they could not produce enough artifact-free tests for all conditions during the pre-processing stage. Before the experiment was conducted, all procedures were approved by the Ethics Committee of the University of Sheffield in the School of Electrical and Electronic Engineering.

Again, the relatively small number of participants included in this study can be attributed to the impact of the COVID pandemic during the data collection period. Restrictions on in-person research and reduced participant availability limited the scope of recruitment, thereby affecting the overall sample size.

4.2.2 The Claw Task

The experimental task employed in this study was adapted from the “Claw Task” introduced in Chapter 3. In the earlier experiment, the task was used exclusively to investigate behavioral changes in trust as a function of machine errors, without the inclusion of EEG recordings. In contrast, the present experiment extends this paradigm by incorporating EEG measurements to investigate the neural correlates of human trust while participants observed machine actions.

The Claw Task can be described as an error-driven brain-computer interface (BCI) scenario in which participants observe machine actions and report their confidence in the system’s ability to successfully complete a given assignment [100]. Specifically, participants observed a computer-simulated arcade “claw crane” game displayed on a monitor (Figure 3.1). The display consisted of eight colored balls aligned horizontally, with a virtual robotic arm positioned above them. At the beginning of each trial, one ball was randomly designated as the target with a preset probability and was colored blue. The remaining seven balls were colored red. The red ball immediately adjacent to the target location was marked as the “on-target neighbor.”

At the start of each run, the claw was positioned directly above one of the red balls, located either two or three positions away from the blue target ball. Approximately once per second, the claw executed one of three possible actions: (1) move one step toward the target ball, (2) move one step away from the target ball, or (3) extend downward to grasp the ball beneath it. Actions were executed instantaneously, with their probabilities determined by whether the claw was aligned with the target or not. When aligned above the blue target ball, the claw could either grasp the ball or step away from it. When misaligned, the claw could move one step closer to the target, one step further away, or grasp the red ball immediately beneath

it. Each run ended after the claw executed a grasp action, after which the screen went blank.

Participants were prompted to report their level of trust in the claw once every minute by pressing a key on a scale of 1–5, with values corresponding to low trust to high trust levels. They were instructed to minimize head and eye movements during the runs but were permitted one-minute breaks between blocks, initiated via an on-screen prompt. Each participant completed eight blocks, each lasting approximately three minutes.

For the purposes of analysis, only erroneous claw actions were considered. Three types of errors were defined: MFFT (Move Further From Target): The claw was aligned above a red non-target ball and moved one step further away from the blue target. SOT (Step Off Target): The claw was aligned above the blue target ball but moved away instead of grasping it. WTI (Wrong Target Item): The claw grasped a red non-target ball.

Both MFFT and SOT errors occurred during horizontal navigation along the row of red balls toward the target blue ball, as shown in Figure 3.1. Importantly, the claw always had information distinguishing whether a lateral movement or grasp action had been executed. Due to the limited number of SOT trials collected, this error type was excluded from the subsequent neurophysiological analysis.

4.2.3 EEG Data Acquisition and Experimental Setup

EEG data were acquired using an Enobio 8 5G system, configured with eight wet electrodes. Wet electrodes were chosen over dry electrodes due to their superior signal quality and reduced impedance variability, which is particularly important for capturing low-amplitude event-related potentials such as the error-related potential (ErrP).

Signals were recorded at a sampling rate of 500 Hz. The electrode montage followed the international 10–20 system, with electrodes positioned at frontal (Fz), central (Cz, C3, C4), parietal (Pz), occipital (Oz), and parieto-occipital (PO7, PO8) sites. These positions were selected to optimize the detection of error-related components, such as the error-related negativity (ERN) and error positivity (Pe), which are typically maximal over fronto-central and parietal regions.

During recording, the reference electrode was placed on the earlobe, and a ground electrode was positioned according to the manufacturer's standard configuration. Electrode–skin contact was ensured by applying conductive gel to maintain low impedance throughout the session. This setup provided a balance between spatial coverage of relevant cortical areas

and participant comfort, while maintaining sufficient signal quality for single-trial classification of error-related responses.

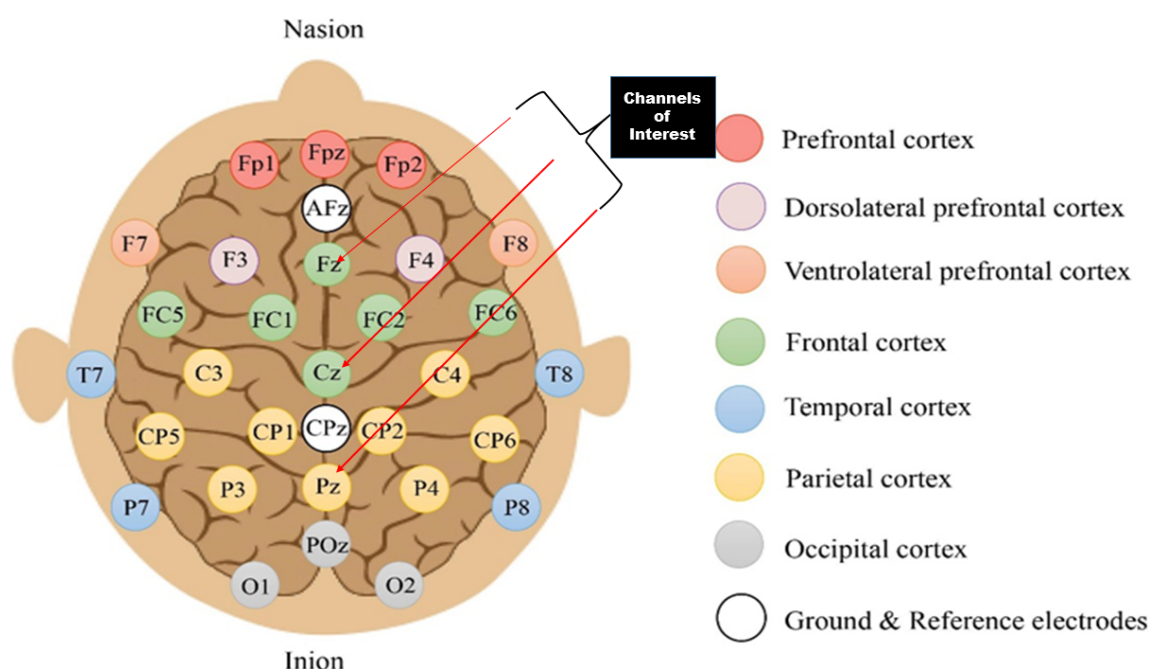


Figure 4.2: Positioning of Electrode within the cortex with specific references to electrode position of interest (Fz, Cz, and Pz)

4.2.4 Preprocessing and Extraction of ErrP Components for Trust Analysis

The recorded EEG signals from eight electrode channels, Frontal-Parietal cortices (Fz, Cz, Pz, Oz, C3, C4, PO7, and PO8), during task performance were resampled to 64 Hz and band-pass filtered between 1 Hz and 10 Hz to remove slow drifts and high-frequency noise. Epochs were extracted from -100 ms to 700 ms relative to the onset of the virtual claw's erroneous movements (MFFT, SOT, and WTI errors) from Fz, Cz, and the Pz electrode, with specific interest in the Pz electrode as evidence from state-of-the-art research indicated that error-related potentials and P300 activities are most prominently observed in the central-midline region of the brain, particularly the Pz electrode [36, 37, 52]. The selection of the time window was also based on visual inspection of averaged time-domain responses, ensuring inclusion of the expected error-related activity (0–100 ms post-stimulus) as shown in Figure 4.7. Each epoch was baseline-corrected to the 0-ms period immediately preceding stimulus onset.

Artifact rejection was conducted automatically by discarding any trials in which the amp-

litude range (difference between the maximum and minimum voltage) in any channel exceeded $-75\mu\text{V}$. This ensured that only clean, artifact-free trials contributed to subsequent averaging and analysis. A full block flow diagram of the processing pipeline is shown in Figure 4.1.

Multiple classification strategies were initially explored to identify potential patterns in participant responses. Notably, the responses from all seven participants fell within only two of the five predefined classification categories, while only two participants provided responses spanning three or more categories. In addition, all participants provided ratings ranging from 1 to 5 on the trust scale. For the purposes of analysis, these responses were grouped into two conditions: ratings of 1–2 were classified as low trust, whereas ratings of 4–5 were classified as high trust. Midpoint ratings (3) were excluded from the analysis. All ratings were obtained using a five-point Likert scale.

Given the limited number of trust-related responses and the uneven distribution across the original classification categories, the robustness of subsequent analyses was inherently constrained. To preserve analytical consistency and enhance interpretability, the classification framework was therefore refined. Categories 1 and 2 were combined to represent low trust, while categories 4 and 5 were merged to represent high trust. The mid-level trust category (Category 3) was excluded from further analysis, as no participant provided responses corresponding to this category.

4.2.5 Time-Domain Analysis of Error-related Potentials (ErrP) across Trust Conditions

The average time-domain error-related potential (ErrP) waveforms were computed for each participant using the preprocessed and extracted EEG trials. To further quantify these components, peak analysis was performed to identify the latencies and amplitudes associated with the error-related negativity (ERN) and Error positivity (Pe). This analysis was based on the grand-averaged time-domain waveform in Pz channel, as shown in Figure 4.7. The ERN was identified as the most prominent negative deflection following the onset of an error, while the Pe was defined as the subsequent positive deflection.

The ERN component was extracted within a time window of 100–200 ms post-stimulus onset, and the Pe within a 300–400 ms window. These temporal boundaries were determined through visual inspection of the grand-average waveforms (Figure 4.7) and aligned with ranges reported in prior literature [36, 52]. Specifically, the ERN window encompassed

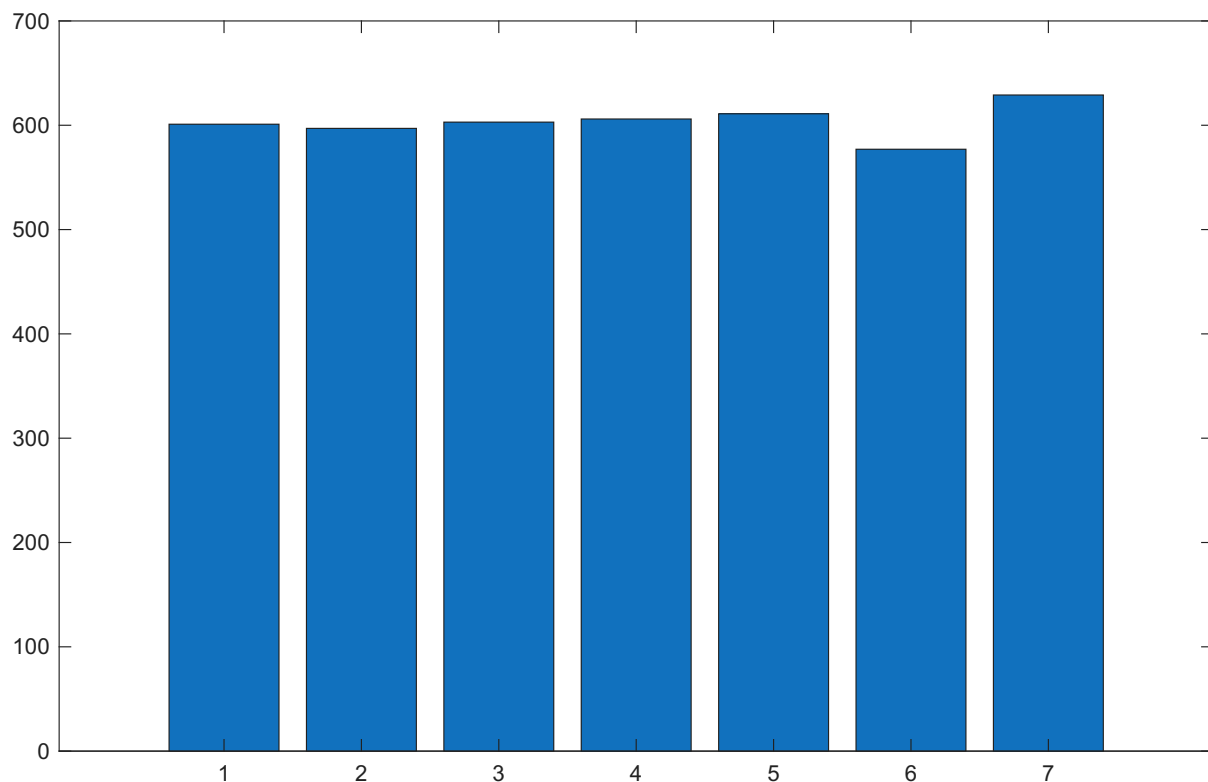


Figure 4.3: Participant-wise distribution of the number of extracted EEG trials recorded during task observation prior to artefact rejection. The y-axis indicates the total number of extracted trials, while the x-axis denotes participant numbers. The distribution highlights individual differences in the quantity of extracted EEG data within the defined time frame. Overall, all seven participants exceeded the chance level of 500 trials per participant.

the interval around the onset and maximum of the negative deflection, whereas the Pe window was centered on the positive peak following the ERN and extended until the return of amplitude values toward baseline. Individual participant averages were examined to extract ERN and Pe peaks within these respective time windows. Average Amplitudes within these extracted time windows for both the ERN and Pe were calculated on the Pz electrode for each participant.

To evaluate whether statistically significant differences in amplitudes existed between the low- and high-trust conditions, a parametric paired-samples t-test was first conducted, followed by a non-parametric Wilcoxon signed-rank test for both the ERN and Pe components. This dual approach was employed to determine whether either test would yield significant statistical outcomes. Specifically, ERN amplitudes and latencies were examined within the 100–200 ms time window, whereas Pe amplitudes were analyzed within the 300–400 ms window.

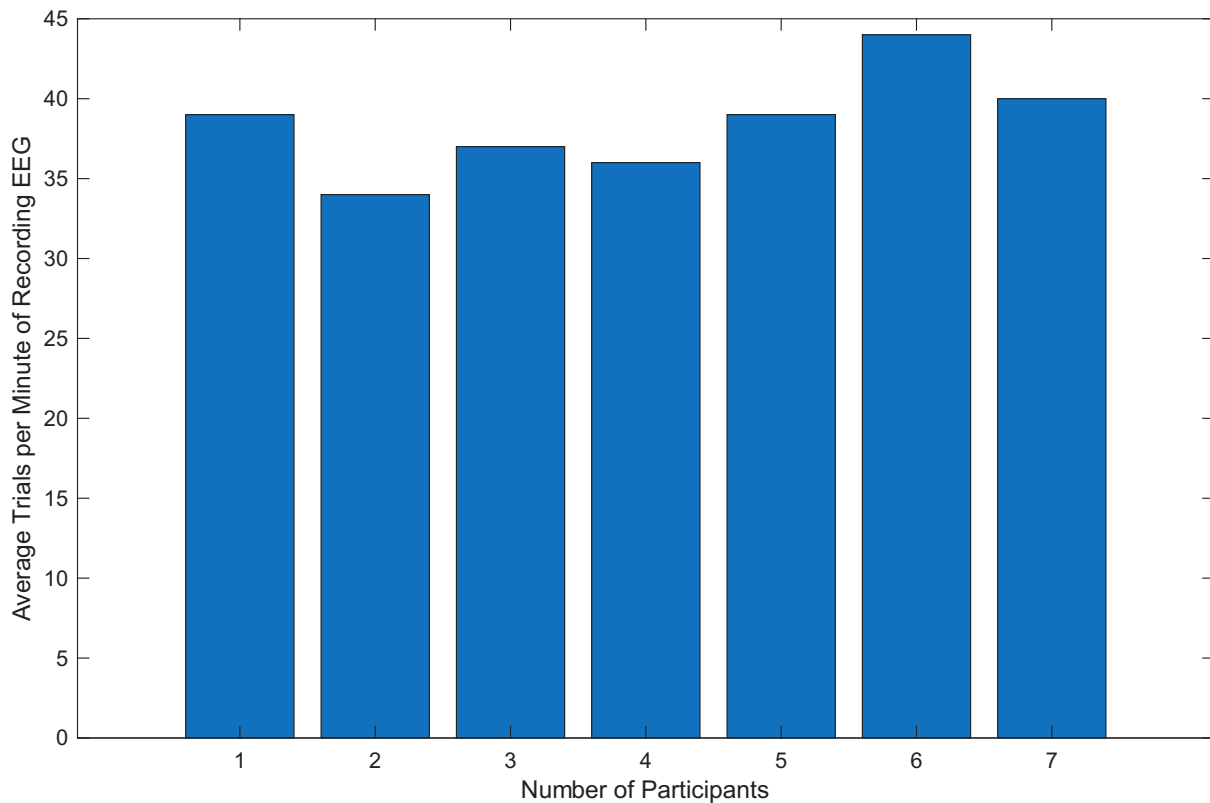


Figure 4.4: Participant-Wise Distribution of Trust Trials Recorded in One-Minute Interval. This figure presents the number of average trust-related trials recorded for each participant during one-minute observation. The distribution reflects individual differences in trust evaluation frequency within the defined time frame

4.2.6 Feature Extraction and Selection for Classification

To identify the most informative subset of extracted features for classification, Stepwise Linear Discriminant Analysis (SWLDA) was employed. The procedure began with an empty model. At each iteration, regression analyses were conducted on models with and without each candidate feature, producing an F-statistic and associated p-value. A threshold of $p < 0.05$ was used as the entry criterion. If at least one feature met this criterion, the feature with the lowest p-value was added to the model. Conversely, if any feature already included in the model exceeded $p > 0.05$ at a given step, the feature with the highest p-value was removed. This iterative process continued until no additional features met the criteria for inclusion or removal. In cases where no features met the entry criterion, the feature with the smallest p-value overall was selected. Feature selection was implemented using the `stepwisefit` function from the MATLAB Statistics and Machine Learning Toolbox (Release 2025b). Both training and test datasets were subsequently reduced to the selected features.

Following feature selection, linear classification models were trained using Linear Discrim-

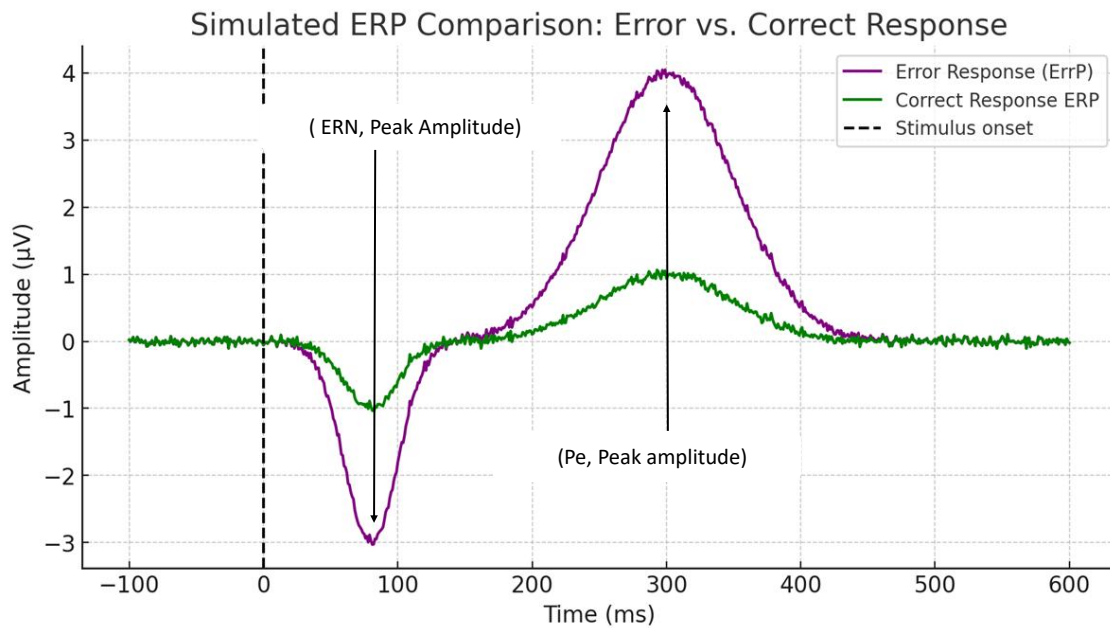


Figure 4.5: Simulated event-related potentials (ERPs) for correct and error responses illustrating the Error-Related Negativity (ERN) and Error Positivity (Pe). The figure shows a simulated comparison of response-locked event-related potentials for error responses (purple) and correct responses (green). Time is aligned to the response at 0 ms, indicated by the dashed vertical line. Following erroneous responses, a pronounced negative deflection appears around 50–100 ms, corresponding to the Error-Related Negativity (ERN), a neural marker associated with rapid performance monitoring processes in the anterior cingulate cortex. This component is followed by a later positive deflection around 200–400 ms, known as the Error Positivity (Pe), which has been linked to conscious error awareness and post-error evaluation. Correct responses show a substantially reduced or absent ERN and Pe. The arrows highlight the approximate peak latencies of these components [37, 101–103].

inant Analysis (LDA) and Support Vector Machines (SVM). Model training was performed using MATLAB’s `fitcdiscr` and `fitsvm` functions, and trial-level predictions were obtained using the `predict` function (MATLAB Statistics and Machine Learning Toolbox, Release 2025b) [61, 104].

4.2.7 Machine Learning-Based Classification of ErrP Features for Trust Level Prediction

Previous studies have proposed several approaches for reporting classification accuracy in related research [52]. One common approach involves reporting the overall accuracy, which represents the proportion of correctly classified trials across all classes. This metric provides a general overview of classifier performance and the relative contribution of each class.

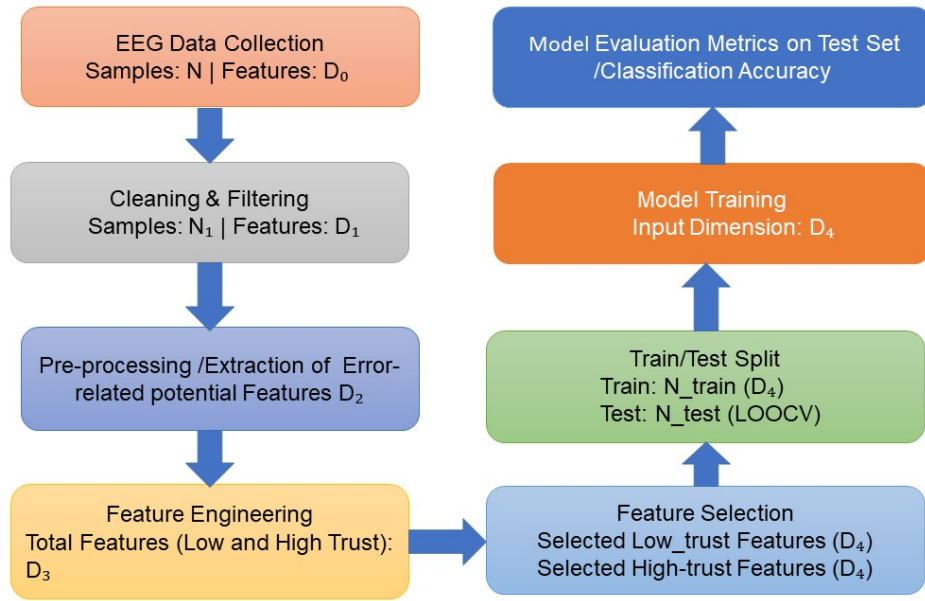


Figure 4.6: Flow Block Pipeline for ErrP feature Extraction and Classification. The dimensionality of the extracted Error-related Potential (ErrP) features for each participant is summarized in Table 4.1. Initially, a pool of approximately 600×3 candidate temporal features was extracted from the three EEG channels, with each channel contributing approximately 20×30 trust time-domain temporal features. Owing to the unequal number of trials recorded under the low-trust and high-trust conditions, the resulting feature space exhibited class imbalance between the two trust categories. To determine the most informative ErrP features for distinguishing between trust levels, Stepwise Linear Discriminant Analysis (SWLDA) was employed as a feature selection method. SWLDA was applied independently to identify representative features associated with both low-trust and high-trust conditions. When multiple features were selected for a particular trust class, their mean value was computed to derive a single representative feature, which was subsequently used as input to the classification models. As reported in Table 4.1, the number of selected features differs across participants, highlighting the presence of inter-subject variability in error-related potential (ErrP) responses.

Table 4.1: The dimensionality of the extracted Error-related Potential (ErrP) features for each participant per channel.

Participant	N (Trials)	Extracted ErrP Features	Low Trust Features	High Trust Features
1	601	38	24	14
2	597	38	30	8
3	603	32	14	18
4	606	35	24	11
5	611	38	18	20
6	577	36	30	6
7	629	27	20	7

Table 4.2: Comparison of Mean Amplitude Differences Between High- and Low-Trust Conditions for the ERN Component Across Participants at the Pz Channel.

Participants	Low Trust Means $\pm$ SDs(μV)	High Trust Means $\pm$ SDs (μV)
1	-0.59 $\pm$ 0.37 μV	0.42 $\pm$ 0.17 μV
2	0.01 $\pm$ 0.31 μV	0.14 $\pm$ 0.13 μV
3	-0.03 $\pm$ 0.10 μV	-0.07 $\pm$ 0.01 μV
4	1.24 $\pm$ 0.27 μV	-1.58 $\pm$ 0.66 μV
5	1.59 $\pm$ 1.10 μV	-0.29 $\pm$ 0.13 μV
6	2.07 $\pm$ 0.23 μV	0.77 $\pm$ 1.77 μV
7	-2.03 $\pm$ 0.31 μV	0.40 $\pm$ 0.13 μV
Grand Average Mean$\pm$S.D	-0.32 $\pm$ 0.24μV	-0.18 $\pm$ 0.42μV

Table 4.3: Summary of Pe Component Mean Amplitudes in High and Low Trust Conditions Across Participants at Pz.

Participants	Low Trust Means $\pm$ SDs(μV)	High Trust Means $\pm$ SDs(μV)
1	0.60 $\pm$ 0.69 μV	0.18 $\pm$ 0.10 μV
2	-0.42 $\pm$ 0.93 μV	-0.16 $\pm$ 0.10 μV
3	0.11 $\pm$ 0.03 μV	-0.07 $\pm$ 0.01 μV
4	1.69 $\pm$ 0.27 μV	-1.92 $\pm$ 0.48 μV
5	1.13 $\pm$ 0.64 μV	-0.17 $\pm$ 0.20 μV
6	4.24 $\pm$ 0.92 μV	2.84 $\pm$ 5.10 μV
7	0.27 $\pm$ 0.09 μV	0.26 $\pm$ 0.11 μV
Grand Average Mean$\pm$S.DV	1.05$\pm$0.51 μV	0.14$\pm$0.87μV

However, it may yield misleading interpretations, particularly when the model exhibits class imbalance or bias.

An alternative approach is the balanced accuracy, where accuracy is computed individually for each class and then averaged across all classes. This method accounts for variations in class performance, offering a more equitable evaluation across categories. Nevertheless, it may obscure certain nuances of classifier behavior by averaging class-specific results.

To classify the ErrP data based on the extracted features associated with participants' reported trust levels, an overall classification approach was implemented using two machine learning algorithms: Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM). These classifiers were selected due to their well-documented effectiveness in EEG signal classification and feature discrimination tasks, as demonstrated in prior research [61, 104]

(Table 2.2).

LDA and SVM are both supervised learning algorithms that have been widely used in brain-computer interface (BCI) research due to their robustness, interpretability, and capacity to handle noisy physiological data. The LDA classifier was particularly suitable for this study because it seeks to maximize the ratio of between-class variance to within-class variance, thereby ensuring optimal class separation in a low-dimensional feature space. This property is advantageous for EEG data, where signal-to-noise ratios are often low and discriminative information is embedded within subtle temporal or spatial patterns.

The SVM classifier was employed as a complementary validation model to evaluate the robustness and generalizability of the LDA results. SVM operates by finding an optimal hyperplane that maximizes the margin between data points from different classes, making it highly effective for complex, non-linearly separable EEG data distributions. Using both classifiers allowed for cross-verification of classification outcomes and minimized the potential for model-specific bias, thereby enhancing the reliability of the trust-level predictions.

Initially, five trust levels were considered for classification, corresponding to the five-point scale used for participants' self-reported trust ratings. However, variability in these ratings resulted in insufficient data within several categories, making it impractical to develop reliable five-class models for all participants. To address this limitation and ensure statistical robustness, the trust ratings were consolidated into two broader categories: low trust and high trust. EEG trials associated with trust ratings of 1–3 were grouped and averaged to represent the low-trust condition, while trials corresponding to ratings of 4–5 were grouped to represent the high-trust condition.

This binary classification approach allowed for a more stable and interpretable analysis of EEG patterns associated with human trust in machines. Furthermore, by training and validating the classification models across these two conditions, the study aimed to establish whether neural features extracted from ErrP data could reliably distinguish between low and high trust states. The combined use of LDA and SVM thus provided a balanced framework—leveraging LDA's computational efficiency and interpretability with SVM's strong discriminative power—to support the detection and prediction of trust-related neural signatures.

An individualized modeling approach was adopted, in which ErrP features extracted from each participant's task responses were used to train and evaluate subject-specific classifiers. This approach ensured that the resulting models captured participant-specific neural signatures associated with trust modulation rather than inter-individual variability.

The LDA classifier was trained using MATLAB's `fitcdiscr` function, while the SVM classifier was trained using the `fitsvm` function, both available in the MATLAB Statistics and Machine Learning Toolbox (version R2025b). These models were trained to distinguish between two categorical trust conditions: Low Trust (ratings 1–3) and High Trust (ratings 4–5), based on the extracted ErrP features.

After model training, predictions were generated for both average and grand-average ErrP trials using MATLAB's `predict` function. Model evaluation was conducted using leave-one-out cross-validation (LOOCV), a rigorous validation technique that iteratively withholds one sample as the test set while training on all remaining samples. This approach provides a nearly unbiased estimate of model generalization performance, particularly beneficial for small or moderate-sized datasets.

Classification performance was evaluated using standard statistical measures of accuracy to quantify the effectiveness of the developed models in distinguishing between the high and low trust conditions. Specifically, the grand average overall classification accuracy was computed across all participants to assess the model's overall ability to correctly identify EEG patterns associated with each trust state.

4.3 Results

4.3.1 Results of Neurophysiological and Statistical Analysis

The results of the paired-samples t-test conducted to compare the amplitudes of the error-related negativity (ERN; 100–200 ms) and the error positivity (Pe; 300–400 ms) between the high- and low-trust conditions revealed no significant differences in amplitude for either component. For the ERN component, the difference between the trust conditions was not statistically significant ($p = 0.60$). Similarly, the Pe component did not show a significant difference between conditions ($p = 0.11$).

The results of the non-parametric Wilcoxon signed-rank test also revealed no statistically significant differences between amplitudes of the trust conditions. Specifically, for the ERN, the test yielded a p-value of 0.57 with a signed-rank value of 18, while for the Pe, a p-value of 0.11 with a signed-rank value of 24 was observed.

These findings align with previous work demonstrating that ERN and Pe amplitudes often show trait-like stability and modest within-subject fluctuations over time and tasks [105, 106].

Specifically, studies have shown that ERN and Pe tend to reflect consistent individual differences rather than being highly sensitive to transient motivational or contextual manipulations. Therefore, while trust is a meaningful cognitive state, its effect on these error-processing components may be smaller than anticipated or mediated by other neural or behavioral factors [107].

As a proof of concept, further analysis was performed to investigate whether any variation occurs between the computed grand average amplitudes of the low and high trust conditions for both the ERN and Pe components within the selected time windows. Interestingly, these results revealed notable differences in grand average amplitude between low and high trust conditions in the Pz channel. For the ERN component, the grand average amplitude under the low trust condition is (Mean= $0.32\mu\text{V}$, SD= $0.24\mu\text{V}$), whereas the high trust condition recorded (Mean= $0.18\mu\text{V}$, SD= $0.42\mu\text{V}$). Similarly, the grand average mean amplitude for the Pe component across all participants is (Mean= $1.05\mu\text{V}$, SD= $0.51\mu\text{V}$) for the low trust condition and (Mean= $0.14\mu\text{V}$, SD= $0.87\mu\text{V}$) for the high trust condition. These results are presented in Tables 4.2 and 4.3, respectively.

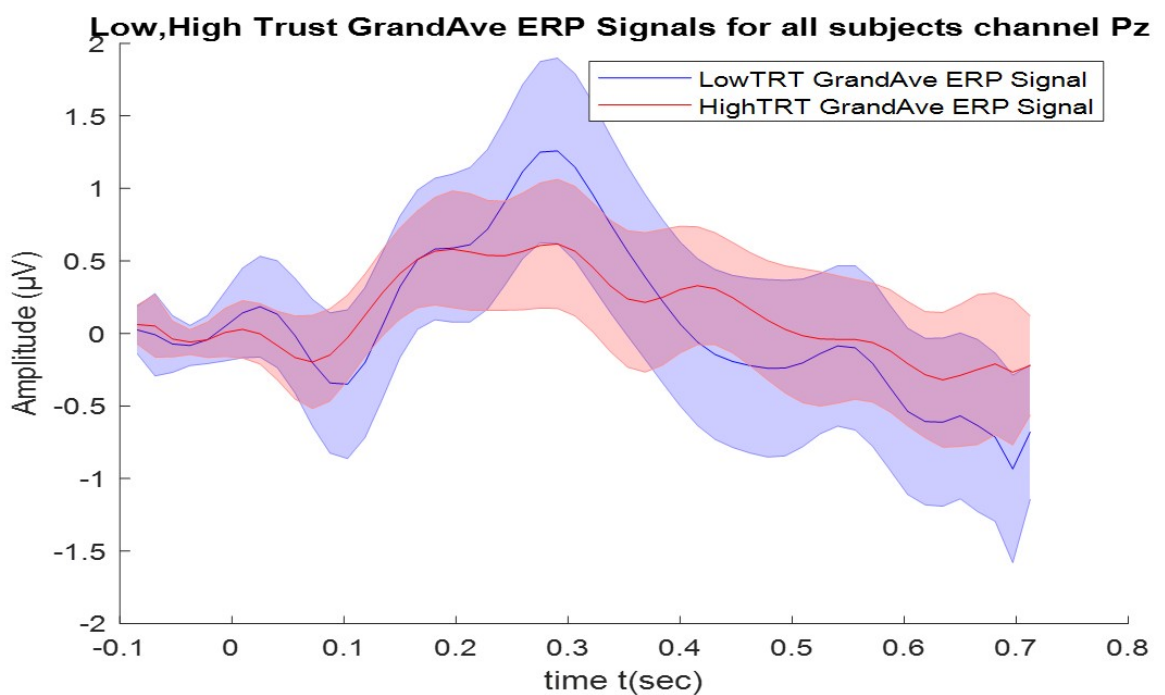


Figure 4.7: Grand-Average Error-Related Potential Amplitudes Across the Pz channels.

Table 4.4: Overall ErrP LDA Classification Accuracies for Individual Participants Using the SWLDA Feature Selection Technique.

Participant	LDA Classification Accuracies(%)	SVM Classification Accuracies(%)
1	83 %	81 %
2	87 %	84 %
3	75 %	48 %
4	78 %	57 %
5	97 %	79 %
6	83 %	80 %
7	52 %	50 %
Mean ± S.D	79 ± 14 %	75 ± 14 %

4.3.2 Classification Performance of LDA and SVM Models

A summary of the classification results obtained from the Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) classifiers is presented in Table 4.4. The results demonstrated that the LDA classifier achieved a grand overall accuracy of $79 \pm 12 \%$, whereas the SVM model produced a slightly lower mean accuracy of $75 \pm 14 \%$. Although this difference is numerically small, it provides meaningful insights into how each model responds to neurophysiological data characteristics in the context of trust-state classification.

The higher mean accuracy obtained by the LDA model suggests that the ErrP features underlying trust-related neural activity were predominantly linearly separable. LDA operates under the assumption that data classes can be distinguished by a linear boundary, which often aligns well with the statistical properties of EEG signals—particularly when the features are extracted from specific frequency bands or time windows associated with cognitive or affective processes [2, 43]. This outcome implies that the neural representations of trust, as captured in the present study, can be effectively modeled using relatively simple linear relationships between features.

In contrast, the SVM classifier, which is generally advantageous for complex, non-linear feature spaces, did not outperform LDA. One possible explanation is that the SVM’s kernel mapping may have introduced unnecessary model complexity, given the linear nature of the underlying ErrP patterns. Moreover, with a limited number of trials for some participants and potential inter-individual variability, the SVM may have been more sensitive to overfitting, thereby reducing its generalization performance. These findings are consistent with previous studies in neurophysiological research, where LDA has frequently demonstrated robust per-

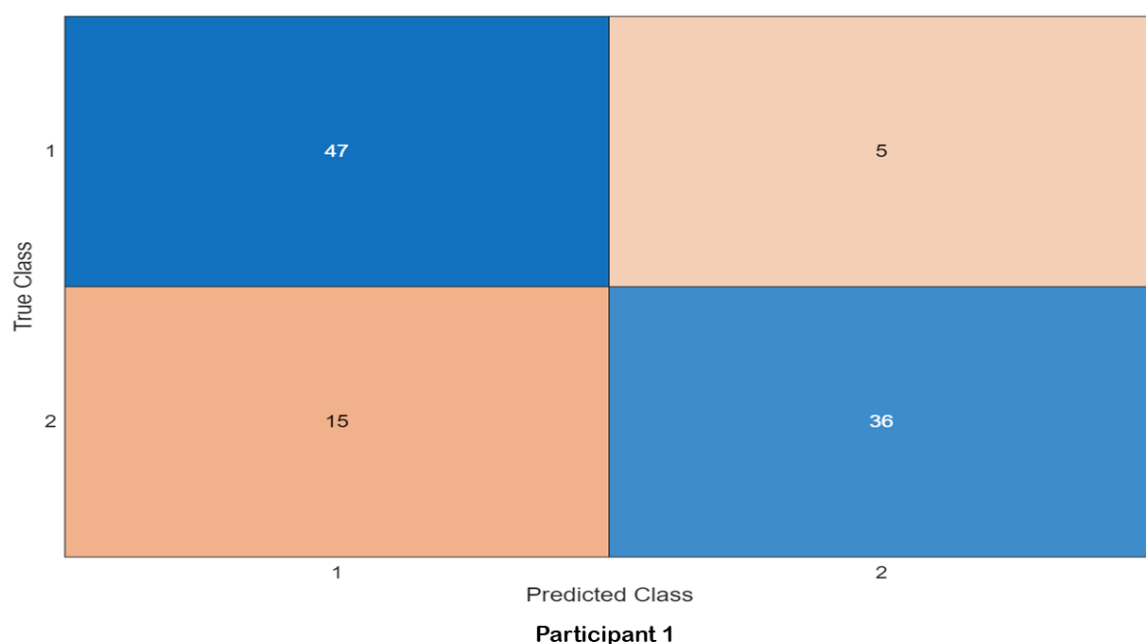


Figure 4.8: Confusion Matrix visualization of Participants 1, LDA performance accuracy

formance in EEG-based classification tasks such as emotion recognition, workload estimation, and decision confidence analysis.

The improved sensitivity of the LDA classifier in identifying neural patterns related to trust states suggests that it was better able to capture subtle fluctuations in cortical activation associated with trust evaluation. Such sensitivity is particularly important in EEG research, where neural signals often exhibit low signal-to-noise ratios and where meaningful patterns may be obscured by inter-trial variability. The LDA's capacity to generalize across participants and channels highlights its suitability for real-time or adaptive trust modeling applications, such as human-robot interaction and neuroadaptive interface design.

Furthermore, to provide a more detailed understanding of the model behavior, confusion matrices were generated individually for some participants (Figures 4.8, 4.9, and 4.10). These matrices allowed visualization of class-wise prediction distributions, highlighting the proportion of correctly and incorrectly classified trials. By examining the diagonal and off-diagonal elements of each confusion matrix, it was possible to identify specific misclassification trends, such as a tendency to overestimate or underestimate a given trust condition. This analysis not only facilitated a clearer interpretation of classifier reliability and robustness across participants but also provided insight into potential sources of variability in EEG-based trust recognition performance.

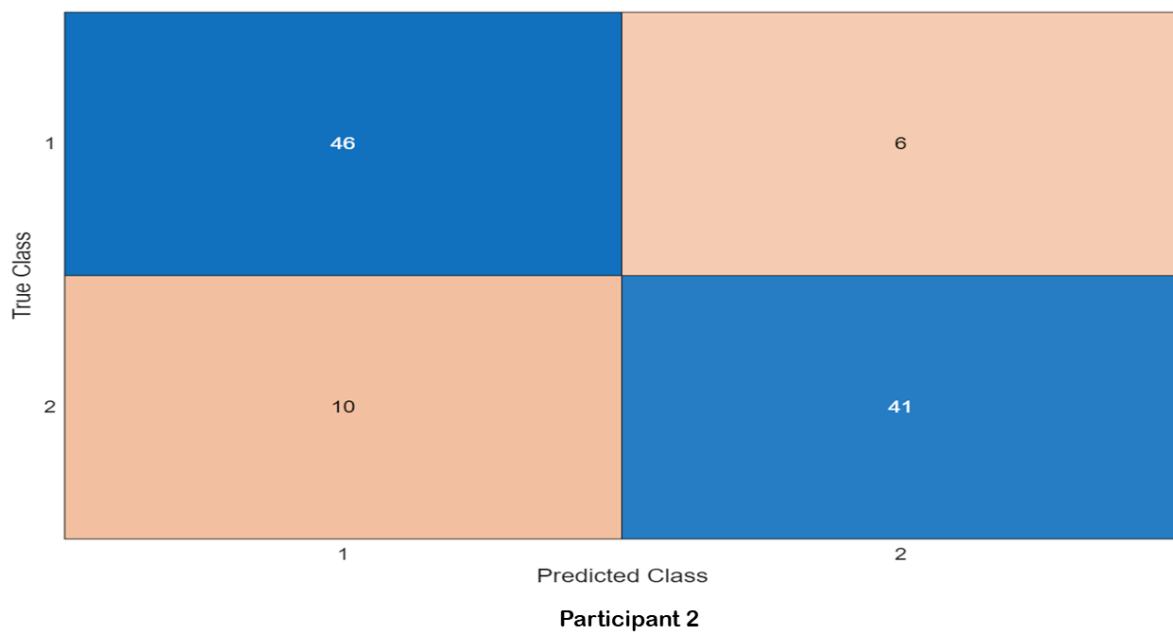


Figure 4.9: Confusion Matrix visualization of Participant 2 performance accuracy

Overall, these results suggest that linear discriminant approaches remain highly effective for EEG-based cognitive state classification, particularly when dealing with moderate sample sizes and features derived from well-defined neurophysiological markers. Future research could explore hybrid approaches that combine the interpretability and efficiency of LDA with the flexibility of non-linear methods such as kernel SVMs or neural networks, potentially leading to improved robustness across diverse cognitive and affective contexts.

4.3.3 Discussion

The paired t-test and Wilcoxon Signed-ranked analysis conducted to compare the amplitudes of the ErrP components (ERN and Pe) between high and low trust conditions did not reveal any significant differences.

As a proof of concept, further analysis was performed to investigate whether any variation occurs between the computed grand average amplitudes of the low and high trust conditions for both the ERN and Pe components within the selected time windows (100-200ms for ERN and 300-400ms for the Pe). Interestingly, these results revealed notable differences in grand average amplitude between low and high trust conditions.

The classification results, as shown in Table 4.4, highlight important distinctions between

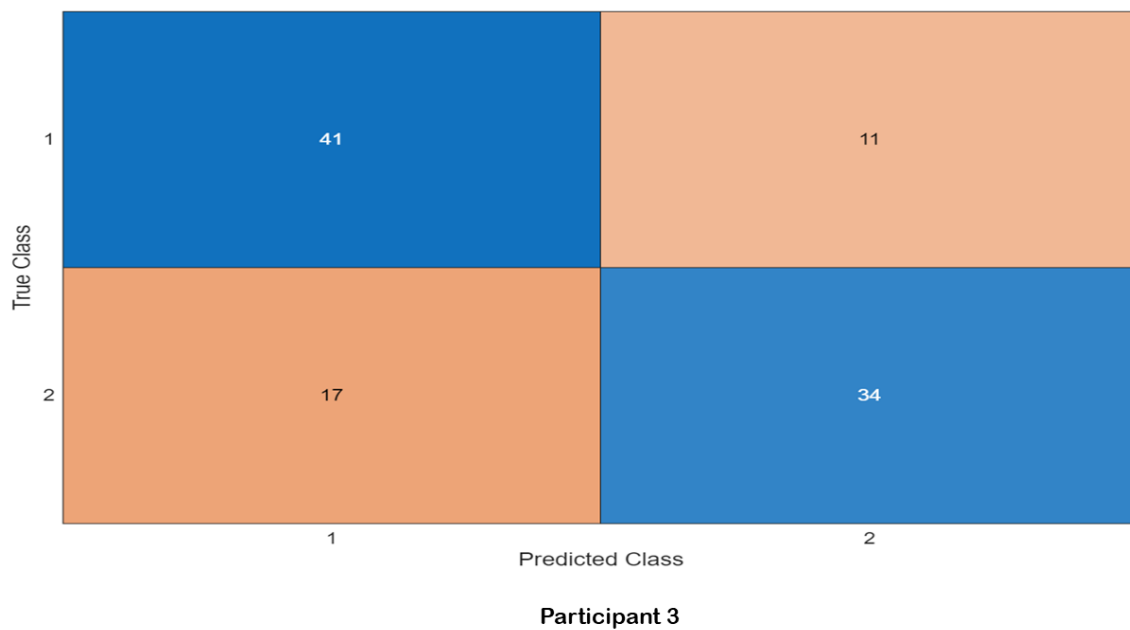


Figure 4.10: Confusion Matrix visualization of Participant 3 performance accuracy

the performance characteristics of linear and non-linear classifiers in the context of EEG-based trust recognition. The superior results obtained with LDA suggest that the EEG feature space was more effectively modeled through linear discriminant structures. LDA's underlying assumption of normally distributed features and its optimization of the ratio between inter-class and intra-class variance enable it to identify linear boundaries that enhance class separability. This makes it particularly advantageous for EEG data characterized by limited dimensionality and moderate noise levels.

In contrast, the SVM classifier demonstrated reduced performance, particularly within the low-trust condition. Although SVMs are theoretically capable of capturing non-linear relationships, their effectiveness in this study may have been limited by factors such as the small number of features, restricted sample size, and the absence of optimized kernel parameters. Moreover, the high degree of variability in EEG signals across participants and trials may have further impacted SVM performance, making it more susceptible to noise and inter-trial inconsistencies without comprehensive parameter tuning.

Taken together, these results suggest that LDA provides a more consistent and interpretable approach for differentiating trust-related EEG patterns under the current experimental framework. The findings align with prior research demonstrating that simpler linear models often outperform complex algorithms in ERP-based classification tasks involving small sample sizes, due to their stability and generalization capabilities. Future investigations should consider in-

corporating feature augmentation, dimensionality reduction, or kernel-based extensions of SVM to better capture potential non-linear relationships and evaluate the scalability of these findings in larger, more heterogeneous datasets.

4.4 Chapter Summary

This chapter presented the recording and analysis of electroencephalogram (EEG) data aimed at investigating the relationship between human trust and machine performance through the use of error-related potentials (ErrPs). Reported trust levels were systematically mapped to the corresponding EEG signals recorded using an 8-channel Enobio (8G) headset. The overarching objective of this study was to explore how variations in human trust influence interaction with semi-autonomous machines, and to determine whether these variations could be reliably detected through ErrP signals.

To achieve this objective, an experimental setup was developed in which a simulated robotic claw navigated through multiple objects, as shown in Figure 3.1. The experimental design was structured to elicit distinct trust-related neural responses by introducing controlled errors during the robot's operation. Specifically, the robotic claw committed three types of errors: (1) moving further away from the target (MFT), (2) stepping off the target (SOT), and (3) wrongly identifying the target (WTI). These error types were chosen to represent distinct forms of machine inaccuracies that could influence participants' perceptions of system reliability and, consequently, their level of trust in the machine.

ErrP signals corresponding to these error events were extracted and analyzed in relation to participants' self-reported trust ratings. Signal preprocessing and analysis were conducted in accordance with the outlined methodologies to ensure consistency and reliability of results. The findings revealed no statistically significant differences in the grand average ErrP amplitudes between high-trust and low-trust conditions, with $p > 0.05$. However, the variation in the computed grand average amplitudes, between low versus high trust conditions, for both ERN and Pe demonstrated a clear difference between low and high trust conditions as shown in tables 4.2 and 4.3.

To evaluate the discriminability of these trust states, a Linear Discriminant Analysis (LDA) classifier was employed to predict trust levels based on the recorded EEG signals. The LDA classifier achieved a grand average classification accuracy of 79% and standard deviation of 14%, demonstrating a strong capability to distinguish between high and low trust conditions. To validate these results, a Support Vector Machine (SVM) classifier was also implemented,

which yielded a comparable classification accuracy of 75% with a standard deviation of 14%. The consistency of results across both classifiers underscores the robustness of the analytical model and the reliability of ErrP-based features in capturing trust-related neural dynamics.

Overall, the results presented in this chapter demonstrate that it is feasible to distinguish between different levels of human trust in semi-autonomous systems by analyzing error-related potentials elicited by the brain in response to machine errors. These findings provide compelling evidence that ErrP-based trust assessment can serve as a viable mechanism for real-time monitoring of human trust in machine agents. In practical applications, such an approach could be utilized to dynamically adapt machine behavior in response to fluctuations in user trust, thereby enhancing human-machine collaboration, improving system safety, and promoting transparency in semi-autonomous environments.

Investigating Human Trust in Machine Using P300 Signals

5.1 Chapter Introduction

In the preceding chapter, it was demonstrated that trust conditions (low versus high trust) can be classified based on the brain's responses to erroneous robot actions, as reflected in error-related potential (ErrP) signals. To maintain methodological rigor and ensure the comparability of findings across experimental contexts, this chapter adopts the same experimental design, data acquisition procedures, preprocessing steps, analytical framework, and classification algorithms as those utilized in the ErrP-based analysis.

As outlined in previous literature, electroencephalography (EEG) has been extensively employed to examine neural responses during the observation of self-initiated and autonomous system movements in navigation-based tasks [36]. Prior studies have shown that EEG signals can effectively discriminate between various types of correct actions [41, 47, 54], enable the subclassification of both erroneous and correct conditions [36, 41, 65], and support reinforcement learning in brain-computer interface (BCI) systems through the classification of correct versus error-related brain activity [41, 47, 54]. These advancements have contributed to the development of intelligent systems capable of autonomously executing low-level

decisions while navigating toward a target, with implicit neural feedback derived from brain activity recorded during task observation [42].

Building upon these advancements in BCI research, the present chapter explores whether human trust in machines can be inferred from EEG responses elicited by correct machine actions. In Chapter 4, trust was examined in relation to the brain's responses to the erroneous movements of a robotic claw performing a factory-like task, in which the claw navigated a designated route to retrieve an object. While that experiment focused on neural responses to task errors that prevented successful completion, it is equally important to investigate how trust manifests when the machine performs its actions accurately and as intended.

Accordingly, this chapter focuses on examining human trust as represented in the brain's responses to correct actions performed by the robotic claw. The analysis is conducted using the same experimental setup and dataset described in Chapter 4, thereby ensuring methodological consistency across all phases of the investigation. The correct actions analyzed include: (i) the claw's movement toward the target (MTT), (ii) the moment at which the target is reached (TR), and (iii) the successful grasp or identification of the correct target (CTI).

5.2 Methods

5.2.1 Participants

Seven participants (mean age = 23 ± 16 years; 4 males and 3 females) took part in this study, and the same individuals also participated in the experiment described in Chapter 4. All participants had normal or corrected-to-normal vision and reported no history of psychiatric conditions, head injury, or photosensitive epilepsy. Written informed consent was obtained prior to participation. All experimental procedures were conducted in accordance with the ethical guidelines approved by the University of Sheffield Ethics Committee, School of Electrical and Electronic Engineering.

As previously mentioned, the relatively small sample size used in this study was largely influenced by the COVID pandemic, which significantly affected recruitment procedures and limited the availability of participants to take part in the experiment.

5.2.2 The Claw Task

The task employed in this section was similar to that described in Chapters 3 and 4. In the present analysis, only correct movements of the claw were considered to have elicited P300 components. These movements included the movements of the simulated robotic claw moved towards the target (MTT), reached the target (TR), and correctly identified the target (CTI), during which the claw successfully grasped the target object.

As previously described, a simulated robotic claw navigated through multiple objects, consisting of a red ball and blue balls. The blue ball indicated the target location, while the red ball represented a non-target location. Participants were instructed to observe the claw's actions and subsequently report their level of trust in the claw's performance. Comprehensive details of the experimental setup and procedure are provided in the preceding section. See Figure 3.1

5.2.3 EEG Acquisition and Experimental Setup

The experimental setup employed in this study was consistent with the configuration described in Chapter 4. Electroencephalographic (EEG) data were recorded using an eight-channel Enobio EEG headset with a sampling rate of 500 Hz. The active electrode sites were positioned according to the international 10–20 system at the following scalp locations: Fz, Cz, Pz, Oz, C3, C4, PO7, and PO8. A reference electrode was placed on the left earlobe to provide a stable reference signal, and all electrode impedances were maintained below manufacturer-recommended thresholds throughout data collection. The classification strategy was implemented using features extracted primarily from Fz, Cz, and the Pz electrodes, based on state-of-the-art studies that ErrP and P300 components reside within the midline central cortex of the brain, particularly on the Pz electrode [36, 37, 52]. However, in this study, Fz was selected for peak analysis as the study demonstrated higher grand average relative peak activities between low and high trust conditions at channel Fz as shown in Figure 5.2.

During the experiment, participants observed the actions of the claw performing object-selection tasks, as described in the preceding sections. EEG signals corresponding to these observation periods were continuously recorded. The raw EEG data were subsequently pre-processed to remove artifacts and segmented into epochs corresponding to specific task events, forming the dataset used for subsequent analyses.

5.2.4 EEG Preprocessing and P300 Component Extraction for Trust Condition Analysis

To identify the most relevant subsets and features of interest for this investigation, the raw EEG data recorded from the Claw Task were preprocessed and analyzed to extract P300 event-related potentials (ERPs) associated with varying levels of participant trust.

The continuous EEG data were first down-sampled to 64 Hz using the `resample` function from the MATLAB Signal Processing Toolbox (version R2025b) to optimize computational efficiency while preserving the temporal resolution necessary for ERP analysis. Subsequently, the data were band-pass filtered between 1 Hz and 10 Hz using a zero-phase Butterworth filter, implemented with the `filtfilt` function (MATLAB Central File Exchange). The use of a zero-phase filter ensured that the phase characteristics of the EEG signals were preserved, maintaining the temporal integrity of the P300 components.

Artifact rejection was conducted automatically by discarding any trials in which the amplitude range (difference between the maximum and minimum voltage) in any channel exceeded $-75\mu\text{V}$. This ensured that only clean, artifact-free trials contributed to subsequent averaging and analysis.

EEG data epochs, ranging from 100ms to 700ms relative to the onset of each correct movement of the claw, and the corresponding levels of reported trust were extracted. These events included the Movement Toward Target (MTT), Target Reached (TR), and Correct Target Identified (CTI) phases. Each epoch was baseline-corrected using the 100 ms pre-stimulus interval (-100 to 0ms) to remove slow drifts and inter-trial variability. During the experiment, participants indicated their trust rating on a five-point Likert scale (1–5), where 1 represented minimum trust and 5 represented maximum trust, by pressing the corresponding numeric key on the keyboard.

For each participant, EEG epochs corresponding to correct claw movements were extracted and categorized according to the self-reported level of trust provided by each trial. For analysis, these trust ratings were collapsed into two categorical conditions: Low Trust (ratings 1–2) and High Trust (ratings 4–5). This processing pipeline allowed for the isolation of clean, temporally aligned P300 responses associated with differential trust conditions. By comparing amplitude and latency characteristics of the P300 component across the low- and high-trust categories, the analysis provided a neurophysiological basis for understanding how human trust modulates during correct robotic actions.

Multiple trials corresponding to each trust condition were pooled together within each participant to compute condition-specific ERPs and further minimize the effects of variability across trials. Each trust condition produced a minimum of 18 trials and at least 35 trials in total per participant. Grand-average ERPs were then generated across participants for each movement type (MTT, TR, CTI) and trust condition (low vs. high).

5.2.5 P300 Component Identification and Statistical Analysis of Trust Conditions

Visual inspection of the grand-average waveforms as shown in Figure 5.2 indicated that the P300 component was most prominent within the 300–400 ms latency window following event onset. Peak analysis was used to identify P300 latencies based on the average and grand average time domain features for the different trust conditions and participants.

Mean amplitudes and standard deviations of all epochs corresponding to low and high trust conditions for each participant were extracted and computed for further analysis. Statistical comparisons between high and low trust conditions were conducted using paired-sample t-tests, followed by a Wilcoxon test. Grand average amplitudes and standard deviation of the extracted features of time domain window of interest (300–400ms) was computed. This analysis aimed to determine whether significant differences in amplitude existed between levels of trust across the different correct movement phases (MTT, TR, and CTI).

5.2.6 Feature Extraction and Selection for Classification

To identify the most pertinent subset features of interest in (Fz,Cz and Pz) for classification analysis, feature selection was carried out using Stepwise Linear Discriminant Analysis (SWLDA), a method well suited for identifying a compact set of predictive features from high-dimensional EEG data. The procedure began with an empty model and iteratively evaluated potential predictors through regression analysis. At each step, candidate features were assessed based on their F-statistics and corresponding p-values. Features meeting the entry threshold ($p < 0.05$) were added, whereas features already in the model that exceeded the removal threshold ($p > 0.05$) were eliminated. This process continued until no additional features met the criteria for inclusion or removal. When no feature satisfied the entry criterion, the predictor with the smallest p-value was selected by default. All feature selection steps were implemented using the `stepwisefit` function in the MATLAB Statistics and Machine Learning Toolbox (Release 2025b), and both training and test datasets were reduced to the final selected feature set.

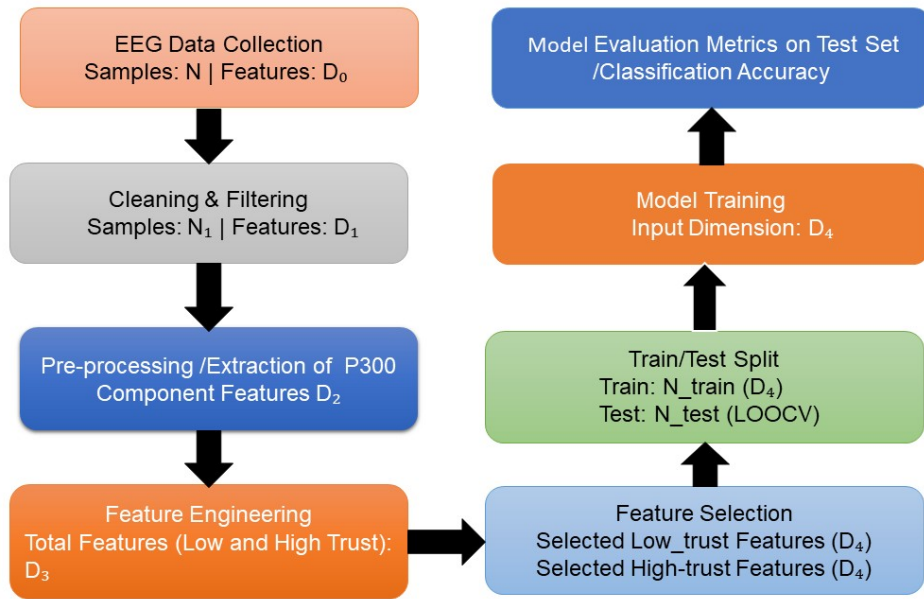


Figure 5.1: Pipeline for P300 Feature Extraction and Classification. The dimensionality of the extracted P300 features for each participant is summarized in Table 4.1. In the initial stage, a set of approximately 600×3 candidate temporal features was derived from three EEG channels. Each channel contributed roughly 20×30 trust time-domain temporal features, forming a high-dimensional feature space. Due to the unequal number of trials recorded under the low-trust and high-trust conditions, the extracted feature set exhibited a degree of class imbalance between the two trust categories. To address the high dimensionality and identify the most discriminative P300 component features for trust-level classification, Stepwise Linear Discriminant Analysis (SWLDA) was employed as the feature selection method. The SWLDA procedure was applied separately to identify representative features associated with both low-trust and high-trust conditions. When multiple features were selected for a particular trust category, their average value was calculated to obtain a single representative feature. This representative value was subsequently used as input to the classification algorithms. As shown in Table 5.1, the number of selected features varies among participants, indicating the presence of inter-subject variability in P300 responses.

Following the selection procedure, classifiers were trained using Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM). Model training was conducted using MATLAB’s `fitcdiscr` and `fitsvm` functions, with trial-level predictions obtained via the `predict` function (MATLAB Statistics and Machine Learning Toolbox, Release 2025b). These approaches are consistent with established EEG-based classification methodologies [61, 104].

Table 5.1: Dimensionality of P300 extracted features for each participant per channel.

Participant	N (Trials)	Extracted P300 Features	Low Trust Features	High Trust Features
1	601	411	235	176
2	597	387	312	95
3	603	343	139	204
4	606	388	262	126
5	611	412	191	221
6	577	400	320	80
7	629	411	325	88

5.2.7 Machine Learning Classification of P300 Features across Trust Conditions

In the preceding chapter, supervised machine learning algorithms—Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM)—were employed to classify error-related potential (ErrP) EEG features across varying trust conditions. In the present chapter, these machine learning algorithms are again utilized as the primary classification strategies for analysing P300 responses under different trust conditions. An individualized modeling approach was adopted, in which EEG features extracted from each participant’s task-related responses were used to train and evaluate subject-specific classifiers. This design ensured that the resulting models captured participant-specific neural signatures associated with trust modulation, thereby minimizing the confounding influence of inter-individual variability.

The LDA classifier was trained using MATLAB’s `fitcdiscr` function, whereas the SVM classifier was trained using the `fitsvm` function, both implemented in MATLAB’s Statistics and Machine Learning Toolbox (version R2025b). The classifiers were trained to distinguish between two categorical trust conditions: Low Trust (ratings 1–2) and High Trust (ratings 4–5), based on features derived from the P300 component.

Following model training, predictions were generated for both average and grand-average EEG trials using MATLAB’s `predict` function. Model validation was performed using LOOCV, a

robust resampling technique in which a single observation is iteratively held out as a test case while the remaining samples are used for training. This method provides an almost unbiased estimate of model generalization, making it particularly suitable for small or moderate-sized datasets.

The implementation of both LDA and SVM classifiers enabled a comparative analysis between linear and non-linear classification strategies. The LDA model, a parametric linear method, maximizes the separation between classes by projecting data along axes that enhance between-class variance while minimizing within-class variance. Conversely, the SVM algorithm identifies an optimal hyperplane with a maximum margin between classes and, through the use of kernel functions, is capable of modeling non-linear decision boundaries. This dual-model approach enhances analytical robustness by verifying classification consistency across distinct algorithmic paradigms.

Classifier performance was quantified using standard evaluation metrics, including accuracy. Additionally, confusion matrices were generated for each participant to visualize prediction distributions and misclassification patterns. Together, these metrics provided a comprehensive evaluation of classification performance across trust conditions and enabled objective comparison between the LDA and SVM models.

5.3 Results

Table 5.2: Overall P300 LDA Classification Accuracies for Individual Participants Using the SWLDA Feature Selection Technique.

Participant	LDA Classification Accuracies(%)	SVM Classification Accuracies(%)
1	82 %	82 %
2	76 %	72 %
3	60 %	50 %
4	75 %	73 %
5	82 %	82 %
6	74 %	77 %
7	47 %	51 %
Mean ± S.D	71 ± 12 %	70 ± 14 %

5.3.1 Neurophysiological and Statistical Analysis

The results obtained from the paired-sample t-test to compare the amplitudes of the P300 components between the high and low trust conditions, extracted between 300-400ms indicated no statistically significant difference in P300 amplitude between the two trust conditions ($p = 0.31$).

To address the non parametric nature of data distribution, a non-parametric Wilcoxon signed-rank test was employed to compare the amplitudes of the P300 components between the high and low trust conditions. The results of the test revealed no statistically significant differences between trust conditions. The test yielded a p-value of 0.38 with a signed-rank value of 20.

Interestingly, analysis of the computed grand average amplitudes for the P300 components within the selected time windows of (300–400 ms), as shown in Figure 5.2, revealed notable differences in mean values between the low and high trust conditions. A grand average (mean = $2.5\mu\text{V}$, SD = $0.10\mu\text{V}$) was recorded under high trust conditions compared to (mean = $5.0\mu\text{V}$, SD = $2.5\mu\text{V}$) under low trust conditions.

5.3.2 Classification Using LDA and SVM

The classification performance of the Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) models was evaluated for differentiating between high- and low-trust conditions based on P300 EEG features. The mean overall classification accuracy of the classifier was $71 \pm 12\%$ using the LDA classifier and $70 \pm 14\%$ using the SVM classifier.

These results suggest that the LDA classifier demonstrated superior discriminative capability in distinguishing between trust conditions. In contrast, the SVM classifier did not outperform LDA. As earlier stated in the preceding chapter, the possible explanation is that the SVM's kernel mapping may have introduced unnecessary model complexity, given the linear nature of the underlying EEG patterns.

A comprehensive summary of the classification results, including the mean accuracies and standard deviations for both the Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) classifiers, is presented in Table 5.2. The application of these classifiers to the EEG data, particularly to the P300 component, provides valuable insight into the variability of neural responses associated with trust-related processing. Consequently, these findings

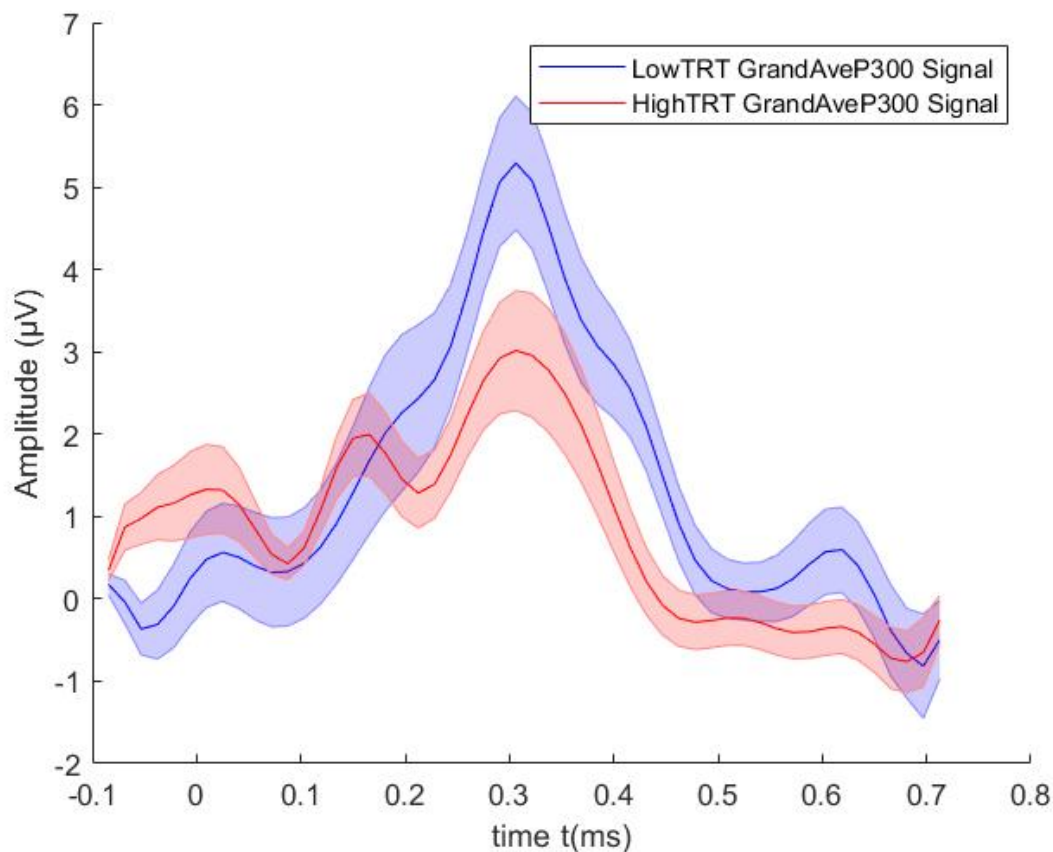


Figure 5.2: Grand-Average P300 Amplitudes at the Fz Channels for all Participants.

highlight the potential of machine learning approaches, such as LDA and SVM, to identify neural markers that can contribute to the modulation and assessment of human trust in human-machine interaction contexts.

Further validation of the LDA classifiers using confusion matrix analyses across participants as shown in Figures 5.3- 5.5 provided additional insight into the classification behavior of the model. The LDA confusion matrices revealed fewer misclassifications between trust conditions compared to SVM, further confirming its relatively stronger discriminative performance under the present experimental conditions.

5.3.3 Discussions

The statistical analysis conducted to compare the amplitudes of the P300 components between high and low trust conditions, using a paired t-test, did not reveal any significant differences (p -value = 0.31), suggesting that variations in trust level did not produce measurable changes

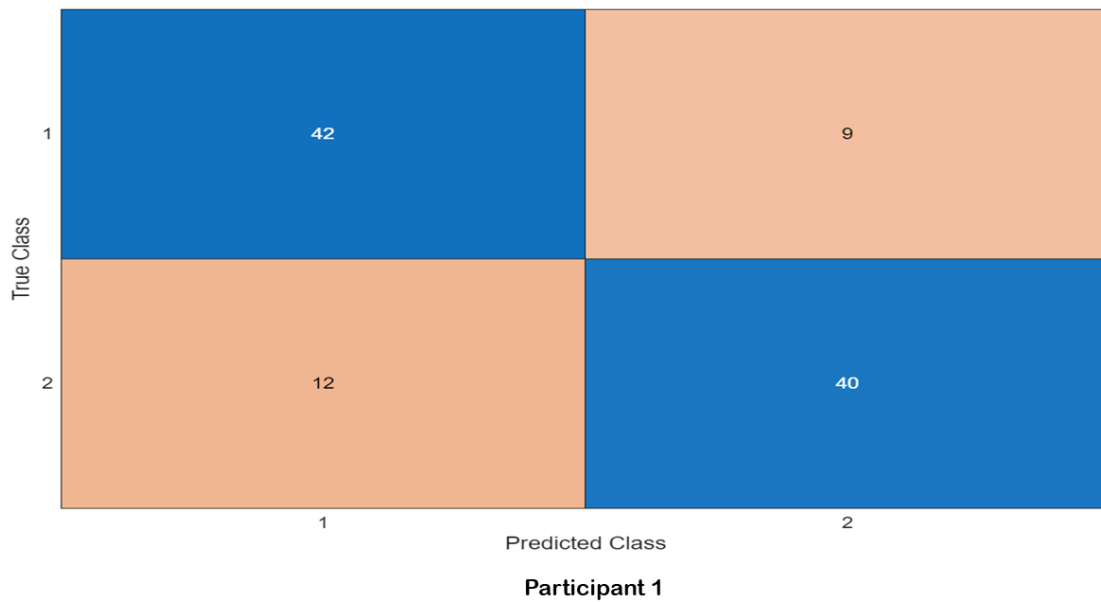


Figure 5.3: Confusion Matrix visualization of Participant 1 performance accuracy

in the P300 response within the examined time window (300–400 ms). To ensure the robustness of these findings, a non-parametric Wilcoxon signed-rank test was performed, which similarly showed no significant differences between trust conditions (p -value = 0.38, signed-rank = 20).

To address these limitations, further analysis of the grand average waveforms revealed observable variations in amplitude between low and high trust conditions. Specifically, the amplitude for the P300 component was higher in the low trust condition (Mean= $5.0\mu\text{V}$, SD= $2.5\mu\text{V}$) compared to the high trust condition (Mean= $2.5\mu\text{V}$, SD= $0.10\mu\text{V}$). Although these differences did not reach statistical significance, the direction of the effect suggests potential modulation of the amplitude by trust level.

Comparing the classification outcomes of the Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) as shown in Table 5.2 suggests that trust-related P300 patterns contain both linearly separable and mildly non-linear components. LDA's performance indicates that much of the discriminative structure in the data can be captured through linear boundaries, while the slightly lower SVM performance supports the presence of fewer non-linear relationships that may not require more flexible decision boundaries.

Importantly, the adoption of participant-specific models proved critical for achieving meaningful classification. The relatively large standard deviations reflect substantial inter-individual

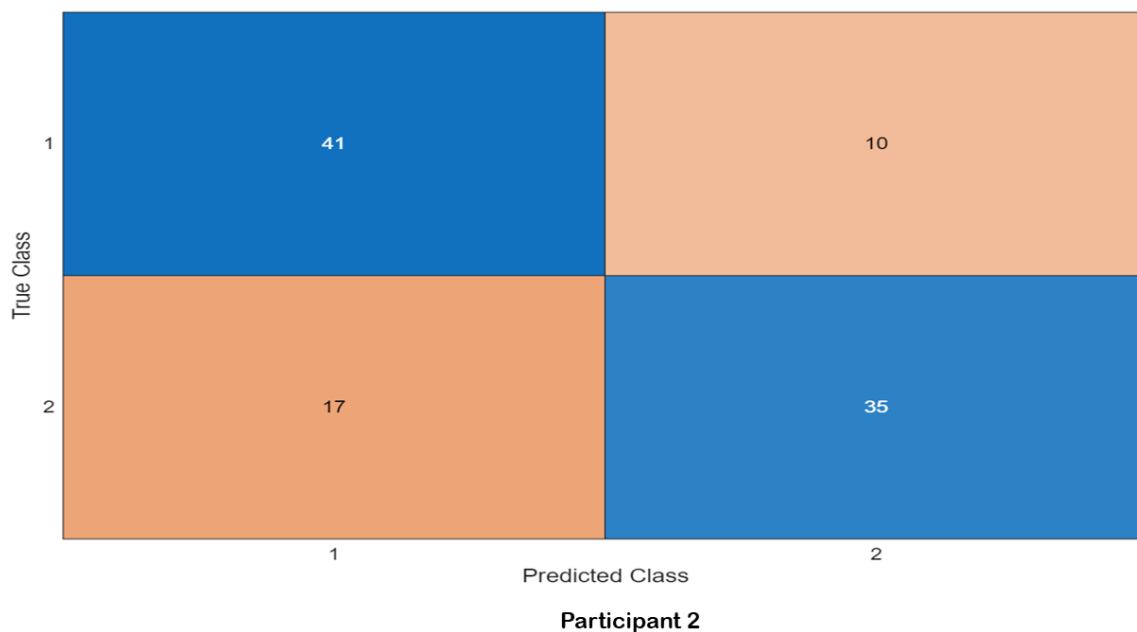


Figure 5.4: Confusion Matrix visualization of Participant 2 performance accuracy

variability in P300 amplitude, latency, and morphology—well-documented characteristics of ERP-based measures. By tailoring each model to individual neural patterns, the classifiers were able to detect subtle, idiosyncratic variations associated with trust modulation that would likely be obscured in a generalized, cross-participant approach.

Overall, these results highlight the feasibility of using ERP-derived features for trust prediction and underscore the importance of individualized modeling strategies in cognitive and affective state classification. While performance levels leave room for improvement, the findings provide a solid foundation for future work exploring more advanced feature extraction methods, adaptive classifiers, and richer neurophysiological markers to enhance the robustness and generalizability of trust detection systems.

5.4 Chapter Summary

This chapter examined the P300 component of EEG signals elicited while participants observed a simulated robotic claw performing three types of correct movements: Movement Toward Target (MTT), Target Reaching (TR), and Correct Target Interaction (CTI). These movement stages provided a framework to investigate neural responses associated with varying levels of trust in the machine.

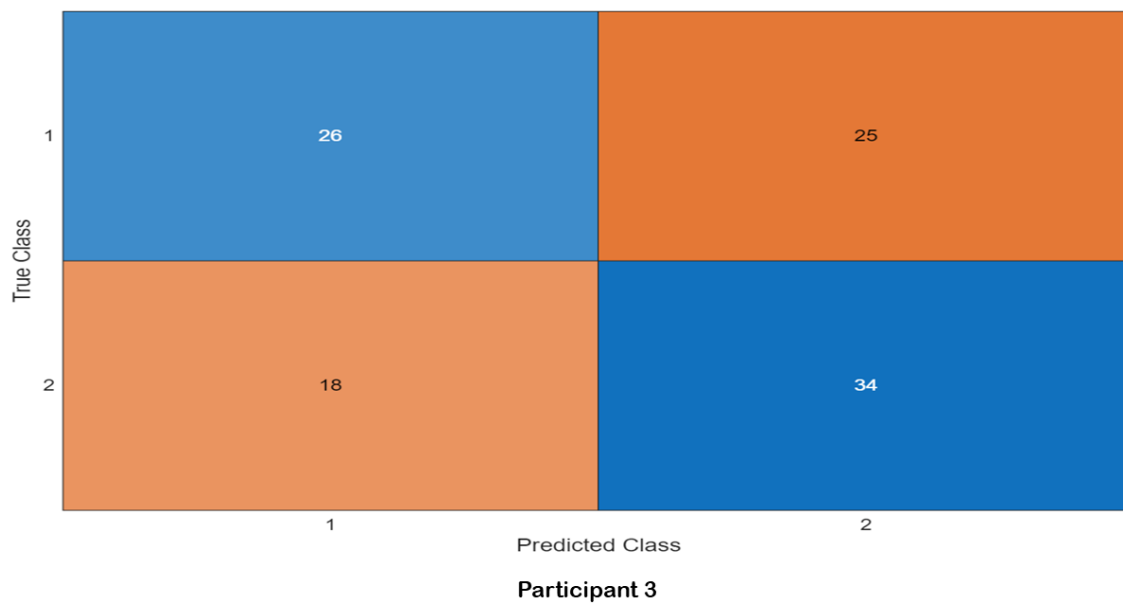


Figure 5.5: Confusion Matrix visualization of Participant 3 performance accuracy

Participant-specific P300 features were extracted and analyzed using Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) classifiers to predict trust levels. The LDA model achieved an overall accuracy of $71 \pm 12\%$, while the SVM model achieved $70 \pm 14\%$, respectively. These results indicate that both linear and non-linear classifiers effectively captured trust-related neural signatures, with individualized modeling enabling the detection of subtle participant-specific variations.

The findings highlight the P300 as a reliable neurophysiological marker of human trust during simulated robotic claw observation and suggest asymmetrical neural processing between low- and high-trust conditions. Importantly, the results have practical implications for human-robot interaction: EEG-based monitoring of trust could inform adaptive and trust-aware robotic systems, allowing real-time adjustments to machine behavior in response to user confidence or uncertainty. This chapter thus establishes a methodological foundation for integrating neural measures into the design of responsive, user-centered robotic interfaces.

General Discussions, Future works and Conclusions

6.1 General Discussions

Generally, State-of-the-art studies [2, 20, 36, 38, 43, 52] have demonstrated that neurocognitive responses associated with error-related potentials (ERPs) and P300 components may reflect emotional and cognitive states such as frustration, disappointment, and attentional engagement. These findings also suggest that such neural signatures are potentially correlated with psychological factors relevant to the assessment of human trust. To integrate behavioral metrics consistent with these insights, the experimental paradigm was designed to allow participants to observe the simulated robotic claw and report their trust levels at approximately one-minute intervals. my analysis focused on EEG signals recorded during these trust-rating periods, specifically those corresponding to the mapped trust categories (see Figure 4.4). Prior to the main experiment, participants completed a 10-minute practice session to become familiar with the task and were instructed to follow all prompts presented on the computer screen. All participants were recruited from similar academic backgrounds and reported comparable levels of experience relevant to the experimental task.

The results presented in Chapter 3 provide the foundational basis for the findings developed

further in Chapters 4 and 5. Collectively, these findings demonstrate that robot performance errors—and, critically, their perceived severities—play a central role in shaping human trust during human–robot interaction (HRI). These observations align with prior research suggesting that trust in automation is highly sensitive to both error type and user interpretation of error significance [2, 20, 36, 38, 43, 52].

A key outcome of this study is the recognition that error severity is not only an objective system attribute but also a subjective human construct. Participants did not uniformly interpret the same error as major or minor; instead, individual judgments varied, consistent with findings in subjective risk perception [2, 20, 43, 52]. As shown in Chapter 3, such subjective assessments significantly affected trust-related behaviours.

Behavioural metrics demonstrated a clear relationship between error severity and trust responses. High-severity errors, specifically the Missing Action (MA) and Wrong Target Identified (WTI) errors, elicited markedly stronger negative trust responses than low severity errors such as the Stepped Off Target (SOT) error. This pattern is consistent with literature showing that severe performance failures reduce trust more rapidly than minor deviations [2, 20, 36, 38, 43, 52].

Within the speed-accuracy trade-off task, trust fluctuations were largely attributable to the types of errors committed by the robot. The results indicated that participants' trust declined most sharply under conditions where the robot exhibited frequent high severity errors. This trend persisted consistently across Chapters 4 and 5, supporting the broader conclusion that error characteristics dominate trust evaluations more strongly than other task variables [2, 97].

Low-trust conditions were associated with increased exposure to severe errors. This, in turn, appeared to elevate participants' cognitive workload, likely due to heightened supervisory demands or vigilance. High trust conditions, by contrast, corresponded to fewer severe errors, reducing cognitive demands and reinforcing perceptions of reliable robot performance.

The electroencephalography (EEG) findings provide convergent evidence for the behavioural observations. Specifically, the pronounced error-related negativity (ERN) peak in the low-trust condition indicates greater cognitive workload and stronger affective responses to perceived robot failures. Elevated ERN amplitudes have been widely associated with increased conflict monitoring, error awareness, and affective responses such as frustration.

These results support the interpretation that frequent severe errors do not merely diminish trust but also increase mental effort and emotional strain in users—a pattern consistent with

cognitive control and workload literature [52].

The combined behavioural and neural findings highlight the usefulness of EEG-based metrics for evaluating and modulating trust in HRI systems. In particular, the P300 component emerges as a promising biomarker for trust modulation. Its sensitivity to outcome relevance and attentional allocation [108], positions it as a valuable neural indicator for adaptive robotic systems capable of responding to human trust states in real time.

These findings also align with prior research [20, 36, 43, 52], suggesting that lower levels of trust heighten performance monitoring and cognitive control. Collectively, these ERP results demonstrate that lower trust in robotic systems intensifies cognitive and affective processing, as evidenced by more pronounced ERN, Pe, and P300 components. These modulations occurred at canonical electrode sites and within expected temporal windows, further confirming their neural specificity. The results support the hypothesis that trust acts as a top-down modulator of attentional and performance-monitoring systems during human-robot interaction.

These neurophysiological responses also suggest engagement of cortical networks involved in error detection and cognitive control, particularly the anterior cingulate cortex (ACC) and posterior parietal cortex, which are commonly associated with the ERN and P300 components, respectively. This reinforces the idea that diminished trust recruits compensatory mechanisms to maintain task performance under uncertainty.

6.2 Limitation

Several limitations were encountered during the course of this study.

- Small sample size, which may limit the statistical power and generalisability of the EEG-related findings.
- Use of simulated tasks instead of physical robot interactions, reducing ecological validity and potentially limiting real-world applicability.
- Restricted EEG electrode configuration, which may constrain spatial resolution and the ability to capture broader neural activity patterns.
- Classification models evaluated only on subject-specific data, limiting insight into cross-subject generalisability and model robustness.

- Extended calibration periods required to capture the full length of recorded EEG signals, increasing setup time and participant fatigue.
- Longer duration for trust-rating procedures, which may contribute to cognitive fatigue and reduce response consistency.

6.3 Future works

While this study has provided significant insights into the neurophysiological mechanisms underlying human trust in semi-autonomous systems, several avenues remain open for further investigation to enhance the robustness, ecological validity, and practical applications of EEG-based trust modeling in human–robot interaction (HRI) [10, 18].

6.3.1 Expansion to Real-World and Multimodal HRI Environments

Future studies should extend the experimental paradigm from a simulated robotic environment to real-world robotic systems involving physical interaction and shared control tasks. Incorporating multimodal physiological signals—such as eye-tracking, galvanic skin response (GSR), heart rate variability (HRV), and facial electromyography (EMG)—alongside EEG could yield richer and more reliable models of human trust. Integrating these modalities may enhance prediction accuracy and support context-aware trust estimation in dynamic, unpredictable environments [109, 110].

6.3.2 Longitudinal and Adaptive Trust Modeling

Trust is inherently dynamic and evolves over time based on experience and performance feedback. Future research could adopt longitudinal designs to examine how neural correlates of trust change during prolonged interaction with autonomous systems. Additionally, adaptive machine learning models, such as reinforcement learning or recurrent neural networks (RNNs), could be employed to track and predict the evolution of trust in real time. This would enable continuous trust calibration, a key requirement for safe and effective collaboration between humans and robots. As demonstrated in the findings of this study, the ErrP and P300 components show strong potential for assessing human trust in machines. However, integrating additional modalities such as EMG or eye-tracking could further enhance the

robustness and reliability of the proposed model [14].

6.3.3 Cross-Individual and Cross-Context Generalization

Although this study employed subject-specific classifiers, developing generalizable models that can adapt to different users without extensive recalibration remains a major challenge. Future work could explore transfer learning, domain adaptation, or federated learning approaches to enhance inter-individual robustness. Moreover, replicating the experiments across diverse task contexts—such as medical robotics, autonomous driving, or military robotics—would validate the broader applicability of ERP-based trust markers [109, 111].

6.3.4 Integration into Closed-Loop Trust-Aware Robotic Systems

A promising direction is the real-time implementation of closed-loop neuroadaptive systems in which robots adjust their autonomy, communication style, or behavior based on the user's inferred trust level. This would require the integration of EEG signal processing pipelines into embedded systems with low-latency feedback loops, allowing robots to modulate their decision-making dynamically. Future work could also evaluate the human factors impact of such adaptive systems, including workload, situation awareness, and perceived transparency [112–115]

6.3.5 Exploration of Additional Neural Markers and Cognitive States

Beyond the ERN, Pe, and P300 components, future research could investigate other EEG signatures such as theta-band synchronization, alpha suppression, or late positive potentials (LPP), which are known to reflect cognitive control, uncertainty, and emotional engagement. Combining ERP and spectral analyses may provide a more holistic understanding of the cognitive-affective dimensions of trust. Additionally, exploring prefrontal and parietal network connectivity using source localization or coherence analysis could yield deeper insights into the distributed neural circuits supporting trust judgments [116–120].

6.3.6 Ethical and Societal Considerations in Neural Trust Monitoring

As trust modeling moves toward real-world deployment, future studies should address ethical, privacy, and interpretability challenges associated with brain-based trust assessment. Investigations into transparent algorithmic decision-making, informed consent, and user autonomy in neuroadaptive systems will be essential to ensure responsible adoption of this technology. Collaboration between neuroscientists, ethicists, and system designers will play a pivotal role in establishing guidelines for the ethical integration of neural data into intelligent systems [121–124].

6.3.7 Multidisciplinary Integration and Cross-Disciplinary Collaboration

Finally, advancing this research area requires strong interdisciplinary collaboration across neuroscience, robotics, artificial intelligence, psychology, and human factors engineering. Future research should aim to develop a unified theoretical framework for trust in HRI, combining cognitive neuroscience evidence with computational trust models to better predict and regulate human–robot collaboration outcomes.

These findings are promising links to bridge the gaps between industrial work environments and neurotechnology adaptive systems that can modulate trust between agents and the machine partners when uncertainties or disappointment occur during system failure. They have opened a research pathway for further investigation as to how trust models can be integrated into real work environments, where the demand for the required work environment to the industries where human and machine collaborative efforts can be completed [11, 14, 18, 47]

6.3.8 Summary

Future research should extend the current findings beyond simulated environments by incorporating real-world human–robot interaction (HRI) scenarios involving physical collaboration and shared control tasks. Expanding the analysis to include multimodal physiological signals—such as eye-tracking, galvanic skin response (GSR), heart rate variability (HRV), and facial electromyography (EMG)—alongside EEG would provide a more comprehensive understanding of trust dynamics and improve model robustness in complex settings. As trust

evolves over time, longitudinal and adaptive modeling approaches using reinforcement learning or recurrent neural networks (RNNs) could enable continuous, real-time trust calibration. To enhance cross-individual and cross-context generalization, future work should explore transfer learning, domain adaptation, and federated learning to develop trust models applicable across users and task domains such as healthcare, autonomous driving, or defense robotics. Integrating these models into closed-loop trust-aware systems could allow robots to adapt their autonomy and communication dynamically based on inferred trust levels, improving collaboration safety and transparency. Further investigations into additional neural markers (e.g., theta synchronization, alpha suppression, and late positive potentials) and functional connectivity analyses could deepen understanding of the cognitive–affective mechanisms underlying trust. Finally, as neural trust monitoring progresses toward practical deployment, it is essential to address ethical, privacy, and transparency concerns through interdisciplinary collaboration among neuroscientists, ethicists, and engineers, ensuring responsible design and implementation of neuroadaptive trust systems that respect user autonomy and societal values.

6.4 Conclusions

This thesis investigated the neurophysiological underpinnings of human trust in machines through a series of behavioural and electrophysiological experiments involving a simulated semi-autonomous robotic claw. Across the three major experimental chapters, the study explored how variations in machine performance, error behaviour, and task accuracy influence human trust, both subjectively (via self-reports) and objectively (via EEG-based neural markers).

6.4.1 Behavioural Analysis of Human Trust Dynamics in Response to Robot Performance and Error Characteristics

In Chapter 3, the behavioural analysis revealed that perceptions of machine reliability and performance differ markedly among individuals. Trust was found to be highly sensitive to variations in error rate, particularly in the case of moving away from target (MA) errors, where increased error frequency led to a significant reduction in reported trust. Conversely, other error types, such as stepping off target (SOT) and wrong target identification (WTI), which occurred at relatively stable rates, did not exhibit significant correlations with trust ratings. Moreover, participants demonstrated a preference for robots operating at lower

speeds but higher accuracy levels, indicating that precision is a stronger determinant of trust than efficiency. Interestingly, when participants were permitted to intervene and correct severe robot errors, their overall trust did not decrease—suggesting that interactive control enhances confidence in the system’s reliability. These findings align with broader evidence on trust in human–robot interaction which emphasises reliability and predictable performance as primary drivers of trustworthiness [11, 18].

6.4.2 Error-related Potentials as Electrophysiological Markers of Trust in Human–Robot Interaction

In Chapter 4, electrophysiological analysis focused on Error-related Potentials (ErrPs) elicited during the observation of erroneous robotic movements—specifically during the phases of movement away from the target, stepping off target, and identifying the wrong target. Results demonstrated that distinct patterns of ErrP activity could differentiate between low and high levels of trust. Enhanced error-related negativity (ERN) and error positivity (Pe) amplitudes were consistently observed under low-trust conditions, reflecting heightened cognitive conflict monitoring and increased conscious error awareness. These findings confirm that error-processing in the human brain dynamically reflects perceived reliability and trust in machine actions, offering a neurophysiological basis for monitoring implicit trust in HRI scenarios. Similar work has shown that brain responses such as feedback-related negativity and P300 vary by perceived trustworthiness of robotic agents [43].

6.4.3 P300 Responses as Neurophysiological Indicators of Trust during Observation of Correct Robotic Actions

In Chapter 5, the investigation extended to correct robotic movements, examining the P300 component as a potential neural indicator of trust during successful task execution. Classification analyses using Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) models achieved mean accuracies of $71 \pm 12\%$ and $70 \pm 14\%$, respectively, in distinguishing between low and high conditions. These results suggest that trust-related differences can be detected even during error-free machine performance, with the P300 serving as a reliable marker of attentional engagement and cognitive evaluation. This is consistent with recent work on EEG-based trust sensing in human–robot interaction [2, 6, 43, 125].

6.4.4 Summary

Taken together, the findings across all experiments establish event-related potentials (ERPs)—specifically the ERN, Pe, and P300 components—as robust and interpretable neurophysiological biomarkers of human trust. Each component captures distinct temporal and cognitive dimensions of trust dynamics: the ERN reflects early conflict monitoring, the Pe indexes conscious error evaluation, and the P300 represents attentional and evaluative processing of machine behaviour. The spatial distribution of these effects— particularly at frontal (Fz) and parietal (Pz) electrode sites—aligns with neural activity in regions such as the anterior cingulate and parietal cortices, which are associated with performance monitoring and decision-making.

Furthermore, by training individualized machine-learning classifiers using ERP features, the study demonstrated the feasibility of real-time neural decoding of trust. A subject-specific LDA classifier achieved an average classification accuracy of 79%, underscoring the discriminative power of EEG-based models in distinguishing trust states. This approach paves the way for the development of adaptive, trust-aware robotic systems that can autonomously adjust behaviour in response to fluctuations in user trust or uncertainty. These ideas echo broader computational modelling efforts in HRI which treat human trust as a latent variable for robot decision-making [10]. Overall, this research contributes novel evidence to the growing field of cognitive neuroscience in human-machine interaction (HMI), confirming that trust is not solely a behavioural or subjective construct but a measurable neural process. The ability to decode trust from brain activity introduces new possibilities for closed-loop HRI systems, where implicit neural cues can inform decision-making, enhance safety, and optimize user experience in high-stakes, precision-dependent environments.

Overall, this thesis validates the use of ERPs as reliable, interpretable, and scalable biomarkers for assessing trust in human-robot collaboration. It further demonstrates that integrating neurophysiological feedback into machine-learning and robotic control frameworks can significantly advance the design of intelligent, adaptive systems—marking a meaningful step toward the realization of seamless, trust-calibrated human-machine partnerships.

Finally, the findings of this study have contributed to the body of knowledge and met the research aims and objectives of this project

- The findings demonstrated behavioural patterns of trust across error rate, error severity, and the human’s ability to intervene to correct erroneous actions of the machine, which is the core objective 1.

- The study was able to substantiate that error-related potential and P300 activities can be used to identify neuro markers that are related to trust.
- The preliminary EEG-based trust model developed in this study to detect low and high trust conditions is an addition to the pool of knowledge within the body of research in the field of neurotechnology and HMLs.

The results of this study were achieved through a systematic progression of the project's objectives. The first objective—investigating how error probability, error rate, and the ability to intervene influence trust—was addressed by developing a simulated robotic claw that mimicked an industrial environment in which the robot executed both correct and erroneous actions. Participants were recruited to observe the robot's behaviour and rate their level of trust, without the use of EEG, while being instructed only to follow the on-screen prompts. Their trust ratings formed the foundational dataset for evaluating the behavioural aspects of trust.

Building on these findings, objectives 2 and 3 focused on identifying EEG responses to erroneous and correct machine actions. To achieve this, a separate set of participants was recruited, with ethical approval from the School of Electrical and Electronic Engineering at the University of Sheffield. Experiments were conducted in the Physiological and Systems Laboratory, where EEG recordings were acquired using an 8-channel wet electrode system as participants repeated the observational task from the first objective. Participants were briefed on the procedure but not the aims of the study, provided informed consent, and completed a short demonstration trial before the main experiment. Rest periods were included between observation blocks to minimise fatigue. Following each session, EEG data were collected for further analysis. These recordings were then processed to extract components of interest—specifically the ERN, Pe, and P300—which were examined in relation to participants' reported trust levels. This enabled the development of a trust-classification model capable of distinguishing between high and low trust states, thereby fulfilling the remaining objectives of the project.

References

- [1] Ewart J. De Visser, Kimberly Goodyear Samuel S. Monfort, Martin O'Hara Li Lu et al. 'A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents'. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39.2 (2017), pp. 116–133 (cit. on pp. 1, 2, 7, 8, 10, 52).
- [2] Eun Soo Jung, Suh Yeon Dong and Soo Young Lee. 'Neural Correlates of Variations in Human Trust in Human-like Machines during Non-reciprocal Interactions'. In: *Scientific Reports* 9.1 (2019) (cit. on pp. 1, 2, 8, 10, 34, 66, 87, 88, 94).
- [3] Simone Erchov. 'Reconceptualizing Trust: Defining, Modeling, and Measuring Trust'. PhD thesis. 2017 (cit. on pp. 2, 9, 33, 34).
- [4] Ann M. Bisantz and Younho Seong. 'Assessment of operator trust in and utilization of automated decision aids under different framing conditions'. In: *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Association, 'Ergonomics for the New Millennium'* 28 (2000) (cit. on pp. 2, 8).
- [5] Ralph Adolphs. *Trust in the brain*. 2002 (cit. on pp. 2, 52).
- [6] Wan Lin Hu, Kumar Akash, Neera Jain and Tahira Reid. 'Real-Time Sensing of Trust in Human-Machine Interactions'. In: *IFAC-PapersOnLine*. Vol. 49. 32. 2016 (cit. on pp. 2, 7, 8, 94).
- [7] J. S. Winston, B. A. Strange, J. O'Doherty and R. J. Dolan. 'Automatic and intentional brain responses during evaluation of trustworthiness of faces'. In: *Nature Neuroscience* 5.3 (2002) (cit. on pp. 2, 10).

- [8] Andrew D. Engell, James V. Haxby and Alexander Todorov. 'Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala'. In: *Journal of Cognitive Neuroscience* 19.9 (2007) (cit. on pp. 2, 10).
- [9] Han Yuan and Bin He. *Brain-computer interfaces using sensorimotor rhythms: Current state and future perspectives*. 2014 (cit. on pp. 2, 11, 13).
- [10] John D. Lee and Katrina A. See. 'Trust in Automation: Designing for Appropriate Reliance'. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46.1 (2004) (cit. on pp. 3, 33, 34, 47, 52, 90, 95).
- [11] John D Lee and Katrina A See. 'Trust in automation: Designing for appropriate reliance Human Factors (In press)'. In: *Human Factors* 46.1 (2004) (cit. on pp. 3, 33, 34, 92, 94).
- [12] John D. Lee. *Review of a pivotal human factors article: Humans and automation: Use, misuse, disuse, abuse*. 2008 (cit. on pp. 3, 33, 34, 52).
- [13] Bies R J. *Trust and Distrust: New Relationships and Realities*. Academy of Management. Tech. rep. 3. 1998 (cit. on pp. 7, 8, 10).
- [14] Kevin Anthony Hoff and Masooda Bashir. 'Trust in automation: Integrating empirical evidence on factors that influence trust'. In: *Human Factors* 57.3 (2015) (cit. on pp. 7, 8, 33, 47, 48, 91, 92).
- [15] Haochen Hu, Kang Yue, Mei Guo, Kai Lu and Yue Liu. 'Subject Separation Network for Reducing Calibration Time of MI-Based BCI'. In: *Brain Sciences* 13.2 (2023) (cit. on p. 8).
- [16] David Cameron, Emily Collins, Hugo Cheung et al. 'Don't worry, we'll get there: Developing robot personalities to maintain user interaction after robot error'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9793. 2016 (cit. on pp. 8, 33).
- [17] David Cameron, Ee Jing Loh, Adriel Chua et al. 'Robot-stated limitations but not intentions promote user assistance'. In: *AISB Annual Convention 2016, AISB 2016*. 2016 (cit. on pp. 8, 33).
- [18] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer et al. 'A meta-analysis of factors affecting trust in human-robot interaction'. In: *Human Factors* 53.5 (2011) (cit. on pp. 8, 48, 90, 92, 94).
- [19] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian and Kerstin Dautenhahn. 'Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust'. In: vol. 2015. June 2015 (cit. on pp. 8, 33, 48).

- [20] Leanne M. Hirshfield, Stuart H. Hirshfield, Samuel Hincks et al. 'Trust in human-computer interactions as measured by frustration, surprise, and workload'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 6780 LNAI. 2011 (cit. on pp. 9, 34, 87–89).
- [21] Murat Kirtay, Erhan Oztop, Anna K. Kuhlen, Minoru Asada and Verena V. Hafner. 'Forming robot trust in heterogeneous agents during a multimodal interactive game'. In: *2022 IEEE International Conference on Development and Learning, ICDL 2022*. 2022 (cit. on p. 9).
- [22] Doreen Jirak, Motonobu Aoki, Takura Yanagi et al. 'Is It Me or the Robot? A Critical Evaluation of Human Affective State Recognition in a Cognitive Task'. In: *Frontiers in Neurorobotics* 16 (2022) (cit. on p. 9).
- [23] Volha Taliaronak, Anna L. Lange, Murat Kirtay, Erhan Oztop and Verena V. Hafner. 'Advancing Humanoid Robots for Social Integration: Evaluating Trustworthiness Through a Social Cognitive Framework'. In: *IEEE International Workshop on Robot and Human Communication, RO-MAN*. 2023 (cit. on p. 9).
- [24] Murat Kirtay, Erhan Oztop, Anna K. Kuhlen, Minoru Asada and Verena V. Hafner. 'Trustworthiness assessment in multimodal human-robot interaction based on cognitive load'. In: *RO-MAN 2022 - 31st IEEE International Conference on Robot and Human Interactive Communication: Social, Asocial, and Antisocial Robots*. 2022 (cit. on p. 9).
- [25] Murat Kirtay, Verena V.V. Hafner, Minoru Asada and Erhan Oztop. 'Trust in Robot-Robot Scaffolding'. In: *IEEE Transactions on Cognitive and Developmental Systems* 15.4 (2023) (cit. on p. 9).
- [26] Michael Novitzky, Paul Robinette, Michael R. Benjamin et al. 'Preliminary Interactions of Human-Robot Trust, Cognitive Load, and Robot Intelligence Levels in a Competitive Game'. In: *ACM/IEEE International Conference on Human-Robot Interaction*. 2018 (cit. on p. 9).
- [27] Michael Novitzky, Paul Robinette, Michael R. Benjamin, Caileigh Fitzgerald and Henrik Schmidt. 'Aquaticus: Publicly Available Datasets from a Marine Human-Robot Teaming Testbed'. In: *ACM/IEEE International Conference on Human-Robot Interaction*. Vol. 2019-March. 2019 (cit. on p. 9).
- [28] Simulink Design Verifier and Apple Hill Drive. 'Simulink[®] Verification and Validation[™] Reference'. In: *ReVision* (2015) (cit. on p. 9).
- [29] Standards Committee. *IEEE Standard for Software Verification and Validation IEEE Standard for Software Verification and Validation*. Vol. 1998. July. 1998 (cit. on p. 9).

- [30] J. S. Winston, B. A. Strange, J. O'doherty and R. J. Dolan. 'Automatic and intentional brain responses during evaluation of trustworthiness of faces'. In: *Social Neuroscience: Key Readings*. 2013 (cit. on p. 10).
- [31] Andrew D Engell, James V Haxby and Alexander Todorov. 'Automatic coding of face properties that signal untrustworthiness in human amygdala'. In: *Magnetic Resonance Imaging* 2002 (2007) (cit. on p. 10).
- [32] Mahnaz Arvaneh and Toshihisa Tanaka. 'Brain-computer interfaces and electroencephalogram: Basics and practical issues'. In: *Signal Processing and Machine Learning for Brain-Machine Interfaces*. 2018 (cit. on pp. 10, 13, 16–18).
- [33] Jianjun Meng, Shuying Zhang, Angeliki Beyko et al. 'Author Correction: Noninvasive Electroencephalogram Based Control of a Robotic Arm for Reach and Grasp Tasks'. In: *Scientific Reports* 10.1 (2020) (cit. on pp. 11, 13, 14).
- [34] Rupert Ortner, Brendan Z. Allison, Gerd Korisek, Herbert Gaggel and Gert Pfurtscheller. 'An SSVEP BCI to control a hand orthosis for persons with tetraplegia'. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 19.1 (2011) (cit. on p. 11).
- [35] Mayur Palankar, Kathryn J. De Laurentis, Redwan Alqasemi et al. 'Control of a 9-DoF wheelchair-mounted robotic arm system using a P300 brain computer interface: Initial experiments'. In: *2008 IEEE International Conference on Robotics and Biomimetics, ROBIO 2008*. 2009 (cit. on p. 11).
- [36] C. Wirth, P. M. Dockree, S. Harty, E. Lacey and M. Arvaneh. 'Towards error categorisation in BCI: Single-trial EEG classification between different errors'. In: *Journal of Neural Engineering*. Vol. 17. 1. 2020 (cit. on pp. 11, 21, 27, 51, 56, 57, 73, 75, 87–89).
- [37] Pierre W. Ferrez and José Del R. Milan. 'You are wrong! - Automatic detection of interaction errors from brain waves'. In: *IJCAI International Joint Conference on Artificial Intelligence*. 2005 (cit. on pp. 11, 21, 25, 27, 29, 56, 60, 75).
- [38] Pierre W. Ferrez and José Del R. Millán. 'Error-related EEG potentials generated during simulated brain-computer interaction'. In: *IEEE Transactions on Biomedical Engineering* 55.3 (2008) (cit. on pp. 11, 21, 27, 29, 87, 88).
- [39] Lucas C. Parra, Clay D. Spence, Adam D. Gerson and Paul Sajda. 'Response error correction - A demonstration of improved human-machine performance using real-time EEG monitoring'. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 11.2 (2003) (cit. on pp. 11, 29).
- [40] Gerwin Schalk, Jonathan R. Wolpaw, Dennis J. McFarland and Gert Pfurtscheller. 'EEG-based communication: Presence of an error potential'. In: *Clinical Neurophysiology* 111.12 (2000) (cit. on p. 11).

- [41] Iñaki Iturrate, Luis Montesano and Javier Minguez. 'Single trial recognition of error-related potentials during observation of robot operation'. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*. 2010 (cit. on pp. 11, 51, 73).
- [42] Iñaki Iturrate, Ricardo Chavarriaga, Luis Montesano, Javier Minguez and José Del R. Millán. 'Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control'. In: *Scientific Reports* 5 (2015) (cit. on pp. 11, 51, 74).
- [43] Ewart J. de Visser, Paul J. Beatty, Justin R. Estep et al. 'Learning From the Slips of Others: Neural Correlates of Trust in Automated Agents'. In: *Frontiers in Human Neuroscience* 12 (2018) (cit. on pp. 14, 33, 66, 87–89, 94).
- [44] Chang S. Nam, Elizabeth Eskander and Sanghyun Choo. 'Neural dynamics of trust in human-robot interaction'. In: *Trust in Human-Robot Interaction*. 2020 (cit. on p. 14).
- [45] 'Reading 15. Automatic and Intentional Brain Responses during Evaluation of Trustworthiness of Faces'. In: *Social Neuroscience*. 2021 (cit. on p. 14).
- [46] T. O. Zander, C. Kothe, S. Welke and M. Roetting. 'Utilizing secondary input from passive brain-computer interfaces for enhancing human-machine interaction'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 5638 LNAI. 2009 (cit. on pp. 14, 15, 51).
- [47] Thorsten O. Zander and Christian Kothe. 'Towards passive brain-computer interfaces: Applying brain-computer interface technology to human-machine systems in general'. In: *Journal of Neural Engineering*. Vol. 8. 2. 2011 (cit. on pp. 14, 15, 51, 73, 92).
- [48] Jens Kohlmorgen, Guido Dornhege, Mikio L. Braun et al. 'Improving Human Performance in a Real Operating Environment through Real-Time Mental Workload Detection'. In: *Toward Brain-Computer Interfacing*. 2019 (cit. on p. 14).
- [49] Laurens R. Krol, Lena M. Andreessen and Thorsten O. Zander. 'Passive Brain-Computer Interfaces: A Perspective on Increased Interactivity'. In: *Brain-Computer Interfaces Handbook: Technological and Theoretical Advances*. 2018 (cit. on p. 15).
- [50] Yuan Pin Lin, Yijun Wang, Chun Shu Wei and Tzyy Ping Jung. 'Human neuroscience assessing the quality of steady-state visual-evoked potentials for moving humans using a mobile electroencephalogram headset'. In: *Frontiers in Human Neuroscience* 8.MAR (2014) (cit. on pp. 15, 34).
- [51] Ali Bashashati, Mehrdad Fatourech, Rabab K. Ward and Gary E. Birch. *A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals*. 2007 (cit. on p. 18).

- [52] Christopher Wirth, Jake Toth and Mahnaz Arvaneh. ““You Have Reached Your Destination”: A Single Trial EEG Classification Study’. In: *Frontiers in Neuroscience* 14 (2020) (cit. on pp. 21, 27, 56, 57, 60, 75, 87–89).
- [53] Martin Spüler and Christian Niethammer. ‘Error-related potentials during continuous feedback: Using EEG to detect errors of different type and severity’. In: *Frontiers in Human Neuroscience* 9.MAR (2015) (cit. on pp. 21, 23, 25, 26).
- [54] Su Kyoung Kim, Elsa Andrea Kirchner, Arne Stefes and Frank Kirchner. ‘Intrinsic interactive reinforcement learning-Using error-related potentials for real world human-robot interaction’. In: *Scientific Reports* 7.1 (2017) (cit. on pp. 21, 25, 47, 51, 73).
- [55] William J Gehring, Michael G H Coles, David E Meyer and Emanuel Donchin. ‘The error-related negativity: an event-related brain potential accompanying errors’. In: *Psychophysiology* 27.4A (1990) (cit. on pp. 21, 22, 51).
- [56] Lüder Deecke, Peter Scheid and Hans H. Kornhuber. ‘Distribution of readiness potential, pre-motion positivity, and motor potential of the human cerebral cortex preceding voluntary finger movements’. In: *Experimental Brain Research* 7.2 (1969) (cit. on p. 21).
- [57] Ziya V. Dikman and John J.B. Allen. ‘Error monitoring during reward and avoidance learning in high- and low-socialized individuals’. In: *Psychophysiology* 37.1 (2000) (cit. on p. 21).
- [58] Fotis P. Kalaganis, Elisavet Chatzilari, Spiros Nikolopoulos, Nikos A. Laskaris and Yian-nis Kompatsiaris. ‘A collaborative representation approach to detecting error-related potentials in SSVEP-BCIs’. In: *Thematic Workshops 2017 - Proceedings of the Thematic Workshops of ACM Multimedia 2017, co-located with MM 2017*. 2017 (cit. on p. 22).
- [59] Blair Kaneshiro, Marcos Perreao Guimaraes, Hyung Suk Kim, Anthony M. Norcia and Patrick Suppes. ‘A representational similarity analysis of the dynamics of object processing using single-trial EEG classification’. In: *PLoS ONE* 10.8 (2015) (cit. on p. 22).
- [60] Lamiaa Abdel-Hamid. ‘An Efficient Machine Learning-Based Emotional Valence Recognition Approach Towards Wearable EEG’. In: *Sensors* 23.3 (2023) (cit. on p. 25).
- [61] E. Donchin, K. M. Spencer and R. Wijesinghe. ‘The mental prosthesis: Assessing the speed of a P300-based brain- computer interface’. In: *IEEE Transactions on Rehabilitation Engineering* 8.2 (2000) (cit. on pp. 25, 60, 62, 79).
- [62] Sandra Rousseau, Christian Jutten and Marco Congedo. ‘The error-related potential and BCIs’. In: *ESANN 2012 proceedings, 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 2012 (cit. on pp. 25, 26, 29, 30).

- [63] Peter S. Bernstein, Marten K. Scheffers and Michael G.H. Coles. "'Where Did I Go Wrong?' A Psychophysiological Analysis of Error Detection'. In: *Journal of Experimental Psychology: Human Perception and Performance* 21.6 (1995) (cit. on pp. 25, 51).
- [64] Akshay Kumar, Lin Gao, Elena Pirogova and Qiang Fang. 'A Review of Error-Related Potential-Based Brain-Computer Interfaces for Motor Impaired People'. In: *IEEE Access* 7 (2019) (cit. on pp. 29, 34).
- [65] G Visconti, B Dal Seno, M Matteucci and L Mainardi. 'Automatic Recognition of Error Potentials in a P300-Based Brain-Computer Interface'. In: *Proceedings of the 4th International Brain-Computer Interface Workshop & Training Course*. (2008) (cit. on pp. 29, 73).
- [66] M. Falkenstein, J. Hohnsbein, J. Hoormann and L. Blanke. 'Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks'. In: *Electroencephalography and Clinical Neurophysiology* 78.6 (1991) (cit. on p. 29).
- [67] Adrien Combaz, Nikolay Chumerin, Nikolay V. Manyakov et al. 'Error-related potential recorded by EEG in the context of a P300 mind speller brain-computer interface'. In: *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2010*. 2010 (cit. on pp. 29, 51).
- [68] Nico M. Schmidt, Benjamin Blankertz and Matthias S. Treder. 'Online detection of error-related potentials boosts the performance of mental typewriters'. In: *BMC Neuroscience* 13.1 (2012) (cit. on p. 29).
- [69] Qingshan She, Tie Chen, Feng Fang et al. 'Improved Domain Adaptation Network Based on Wasserstein Distance for Motor Imagery EEG Classification'. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023) (cit. on p. 29).
- [70] Dan Zhang, Huaying Song, Honglai Xu et al. 'An N200 speller integrating the spatial profile for the detection of the non-control state'. In: *Journal of Neural Engineering* 9.2 (2012) (cit. on p. 29).
- [71] Rozhin Yousefi, Alborz Rezazadeh Sereshkeh and Tom Chau. 'Exploiting error-related potentials in cognitive task based BCI'. In: *Biomedical Physics and Engineering Express* 5.1 (2019) (cit. on p. 29).
- [72] Eric W. Sellers, Dean J. Krusienski, Dennis J. McFarland and Jonathan R. Wolpaw. 'Noninvasive Brain-Computer Interface Research at the Wadsworth Center'. In: *Toward Brain-Computer Interfacing*. 2019 (cit. on p. 30).
- [73] Jeanne M. Wilson, Susan G. Straus and Bill McEvily. 'All in due time: The development of trust in computer-mediated and face-to-face teams'. In: *Organizational Behavior and Human Decision Processes* 99.1 (2006) (cit. on p. 33).

- [74] Raja Parasuraman and Victor Riley. 'Humans and Automation: Use, Misuse, Disuse, Abuse'. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39.2 (June 1997), pp. 230–253 (cit. on pp. 33, 34, 52).
- [75] Raja Parasuraman, Victor Riley, Jason Furman et al. 'Artificial Intelligence , PREPARING FOR THE FUTURE Automation , and the Economy'. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39.1 (2016) (cit. on pp. 33, 34).
- [76] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay and Michael L. Walters. 'The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario'. In: *Paladyn* 9.1 (2018) (cit. on p. 34).
- [77] Peter de Vries, Cees Midden and Don Bouwhuis. 'The effects of errors on system trust, self-confidence, and the allocation of control in route planning'. In: *International Journal of Human Computer Studies* 58.6 (2003) (cit. on p. 34).
- [78] Roger C. Mayer, James H. Davis and F. David Schoorman. 'An Integrative Model Of Organizational Trust'. In: *Academy of Management Review* 20.3 (1995) (cit. on pp. 34, 52).
- [79] Roger C. Mayer, James H. Davis and F. David Schoorman. 'An Integrative Model of Organizational Trust'. In: *The Academy of Management Review* 20.3 (1995) (cit. on p. 34).
- [80] Seeung Oh, Younho Seong, Sun Yi and Sangsung Park. 'Investigation of human trust by identifying stimulated brain regions using electroencephalogram'. In: *ICT Express* 8.3 (2022) (cit. on p. 34).
- [81] Enid N.H. Montague, Brian M. Kleiner and Woodrow W. Winchester. 'Empirically understanding trust in medical technology'. In: *International Journal of Industrial Ergonomics* 39.4 (2009) (cit. on p. 34).
- [82] Aniana Cruz, Gabriel Pires and Urbano J. Nunes. 'Generalization of ErrP-Calibration for Different Error-Rates in P300-Based BCIs'. In: *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*. 2018 (cit. on p. 34).
- [83] Manuel Seet, Jonathan Harvy, Rohit Bose et al. 'Differential Impact of Autonomous Vehicle Malfunctions on Human Trust'. In: *IEEE Transactions on Intelligent Transportation Systems* 23.1 (2022) (cit. on p. 34).
- [84] Huaijian Zhang, Ricardo Chavarriaga, Mohit Kumar Goel, Lucian Gheorghe and Jose Del R. Millan. 'Improved recognition of error related potentials through the use of brain connectivity features'. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. 2012 (cit. on p. 34).

- [85] Rui Zhang, Yadi Chen, Zongxin Xu et al. 'Recognition of single upper limb motor imagery tasks from EEG using multi-branch fusion convolutional neural network'. In: *Frontiers in Neuroscience* 17 (2023) (cit. on p. 34).
- [86] Niki Panteli, Jason R.C. Nurse, Emily Collins and Nikki Williams. 'Trust disruption and preservation in the Covid-19 work from home context'. In: *Journal of Workplace Learning* 35.3 (2023) (cit. on p. 34).
- [87] Andreea Isabela Varga, Ivan Spehar and Helge Skirbekk. 'Trustworthy management in hospital settings: a systematic review'. In: *BMC Health Services Research* 23.1 (2023) (cit. on p. 34).
- [88] Y. Serkan Ozmen. 'How employees define organisational trust: analysing employee trust in organisation'. In: *Journal of Global Responsibility* 9.1 (2018) (cit. on p. 34).
- [89] Maria João Velez. 'The relationship between abusive supervision and organizational trust: The role of subordinates' self-esteem'. In: *Análise Psicológica* 38.2 (2020) (cit. on p. 34).
- [90] Happy Chidi Onyeoruru, Christopher Wirth, Joshua Giles and Mahnaz Arvaneh. 'Analysing and Modelling Human Trust to a Navigation Robot'. In: *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering, MetroXRaine 2023 - Proceedings*. 2023 (cit. on p. 35).
- [91] Hanzhang Lu, Feng Xu, Ksenija Grgac et al. 'Calibration and validation of TRUST MRI for the estimation of cerebral blood oxygenation'. In: *Magnetic Resonance in Medicine* 67.1 (2012) (cit. on p. 48).
- [92] P. Madhavan and D. A. Wiegmann. 'Similarities and differences between human-human and human-automation trust: An integrative review'. In: *Theoretical Issues in Ergonomics Science* 8.4 (2007) (cit. on p. 48).
- [93] Alexandria Meyer. *A biomarker of anxiety in children and adolescents: A review focusing on the error-related negativity (ERN) and anxiety across development*. 2017 (cit. on p. 48).
- [94] Robert Hoffman, Shane Mueller, Gary Klein and Jordan Litman. *Measuring Trust in the XAI Context*. 2021 (cit. on p. 48).
- [95] Mark T. Dzindolet, Hall P. Beck and Linda G. Pierce. 'Network operations: Developing trust in human and computer agents'. In: *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology*. 2010 (cit. on p. 48).

- [96] R. C Mayer, J. H Davis and F. D. Schoorman. 'An integrative model of organizational trust. *Academy of management review*'. In: *Academy of Management Review* 20.3 (1995) (cit. on p. 52).
- [97] Roy J. Lewicki, Daniel J. McAllister and Robert I. Bies. 'Trust and distrust: New relationships and realities'. In: *Academy of Management Review* 23.3 (1998) (cit. on pp. 52, 88).
- [98] R O Y J Lewicki and Robert J Bies. 'Trust and Distrust : New Relationships and Realities'. In: *Academy of Management Review* 23.3 (2019) (cit. on p. 52).
- [99] Roy J. Lewicki, Daniel J. McAllister and Robert J. Bies. 'Trust and Distrust: New Relationships and Realities'. In: *The Academy of Management Review* 23.3 (1998) (cit. on p. 52).
- [100] Erin K. Chiou and John D. Lee. 'Trusting Automation: Designing for Responsivity and Resilience'. In: *Human Factors* 65.1 (2023) (cit. on p. 54).
- [101] Thérèse J.M. Overbeek, Sander Nieuwenhuis and K. Richard Ridderinkhof. *Dissociable components of error processing: On the functional significance of the Pe vis-à-vis the ERN/Ne*. 2005 (cit. on p. 60).
- [102] Francesco Di Gregorio, Martin E. Maier and Marco Steinhauser. 'Are errors detected before they occur? Early error sensations revealed by metacognitive judgments on the timing of error awareness'. In: *Consciousness and Cognition* 77 (2020) (cit. on p. 60).
- [103] Francesco Di Gregorio, Martin E. Maier and Marco Steinhauser. 'Errors can elicit an error positivity in the absence of an error negativity: Evidence for independent systems of human error monitoring'. In: *NeuroImage* 172 (2018) (cit. on p. 60).
- [104] Wenna Chen, Yixing Wang, Yuhao Ren et al. 'An automated detection of epileptic seizures EEG using CNN classifier based on feature fusion with high accuracy'. In: *BMC Medical Informatics and Decision Making* 23.1 (2023) (cit. on pp. 60, 62, 79).
- [105] Michael J. Larson, Scott A. Baldwin, Daniel A. Good and Joseph E. Fair. 'Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): The role of number of trials'. In: *Psychophysiology* 47.6 (2010) (cit. on p. 64).
- [106] Hans Kirschner, Jil Humann, Jan Derrfuss, Claudia Danielmeier and Markus Ullsperger. 'Neural and behavioral traces of error awareness'. In: *Cognitive, Affective and Behavioral Neuroscience* 21.3 (2021) (cit. on p. 64).
- [107] Doreen M. Olvet and Greg Hajcak. 'The stability of error-related brain activity with increasing trials'. In: *Psychophysiology* 46.5 (2009) (cit. on p. 65).

- [108] *Correction to: Passive Elaborative Cueing of Retrieval Processes Following Traumatic Brain Injury (Brain Injury, (2021), 35, 10, (1168-1183), 10.1080/02699052.2021.1953595).* 2024 (cit. on p. 89).
- [109] Kumar Akash, Tahira Reid and Neera Jain. 'Adaptive Probabilistic Classification of Dynamic Processes: A Case Study on Human Trust in Automation'. In: *Proceedings of the American Control Conference*. Vol. 2018-June. 2018 (cit. on pp. 90, 91).
- [110] Shenmao Zhang, Linsheng Zhong, Yang Zou et al. 'Improved multiplier-free Mueller-Müller Baud-rate timing error detector for optical IM/DD system'. In: *Optics Express* 29.26 (2021) (cit. on p. 90).
- [111] Tyler Mari, Oda Asgard, Jessica Henderson et al. 'External validation of binary machine learning models for pain intensity perception classification from EEG in healthy individuals'. In: *Scientific Reports* 13.1 (2023) (cit. on p. 91).
- [112] Wenjie Jin, Xin Xin Zhu, Lifeng Qian et al. *Electroencephalogram-based adaptive closed-loop brain-computer interface in neurorehabilitation: a review*. 2024 (cit. on p. 91).
- [113] Yuchong Zhang, Nona Rajabi, Farzaneh Taleb et al. 'Mind Meets Robots: A Review of EEG-Based Brain-Robot Interaction Systems'. In: *International Journal of Human-Computer Interaction* 41.20 (2025) (cit. on p. 91).
- [114] Rodrigo Vieira, Plinio Moreno and Athanasios Vourvopoulos. 'EEG-based action anticipation in human-robot interaction: a comparative pilot study'. In: *Frontiers in Neurobotics* 18 (2024) (cit. on p. 91).
- [115] Chengguo Liu, Kai Zhao, Weiyong Si, Junyang Li and Chenguang Yang. 'Neuroadaptive Admittance Control for Human-Robot Interaction With Human Motion Intention Estimation and Output Error Constraint'. In: *IEEE Transactions on Cybernetics* 55.6 (2025) (cit. on p. 91).
- [116] Felix Bacigalupo and Steven J. Luck. 'Alpha-band EEG suppression as a neural marker of sustained attentional engagement to conditioned threat stimuli'. In: *Social Cognitive and Affective Neuroscience* 17.12 (2022) (cit. on p. 91).
- [117] Yiwei Li, Sijin Li, Yuyao Tang, Shuai Hao and Dandan Zhang. 'Causal evidence for the role of prefrontal theta oscillations in emotion regulation using neurofeedback training'. In: *NeuroImage* 320 (2025) (cit. on p. 91).
- [118] Stephen B. R. E. Brown, Henk van Steenbergen, Guido P. H. Band, Mischa de Rover and Sander Nieuwenhuis. 'Functional significance of the emotion-related late positive potential'. In: *Frontiers in Human Neuroscience* 6 (2012) (cit. on p. 91).

- [119] Majid Roshanaei, Hamzeh Norouzi, Julie Onton, Scott Makeig and Alireza Mohammadi. 'EEG-based functional and effective connectivity patterns during emotional episodes using graph theoretical analysis'. In: *Scientific Reports* 15.1 (2025) (cit. on p. 91).
- [120] John Polich. *Updating P300: An integrative theory of P3a and P3b*. 2007 (cit. on p. 91).
- [121] Finale Doshi-Velez and Been Kim. 'A Roadmap for a Rigorous Science of Interpretability'. In: *arXiv preprint arXiv:1702.08608v1* (2017) (cit. on p. 92).
- [122] José M. Muñoz. 'Achieving cognitive liberty The Battle for Your Brain: Defending the Right to Think Freely in the Age of Neurotechnology Nita A. Farahany St. Martin's Press, 2023. 288 pp. ' In: *Science* 379.6637 (2023) (cit. on p. 92).
- [123] Marcello Lenca and Roberto Andorno. 'Towards new human rights in the age of neuroscience and neurotechnology'. In: *Life Sciences, Society and Policy* 13.1 (2017) (cit. on p. 92).
- [124] David McMenemy. 'Ethical Issues in Brain-Computer Interfaces'. In: *Intelligent Computing for Interactive System Design*. 2021 (cit. on p. 92).
- [125] Mengyang Xu, Binshi Bo, Mengchao Pei et al. 'High-resolution relaxometry-based calibrated fMRI in murine brain: Metabolic differences between awake and anesthetized states'. In: *Journal of Cerebral Blood Flow and Metabolism* 42.5 (2022) (cit. on p. 94).