RÉPUBLIQUE FRANÇAISE
*Liberté*
*Égalité*
*Fraternité*

ISAE SUPAERO Institut Supérieur de l'Aéronautique et de l'Espace

Université de Toulouse

# Doctorat de l'Université de Toulouse

## délivré par l'ISAE-SUPAERO

---

# Détection d'états cognitifs pour la collaboration humain-robot : une approche électrophysiologique

---

Thèse présentée et soutenue, le 27 novembre 2025 par

**Mathias RIHET**

**École doctorale**
SYSTEMES

**Spécialité**
Systèmes embarqués et Robotique

**Unité de recherche**
ISAE-ONERA DECISIO DECIsion, Supervision et Interaction pour l'Opération de systèmes complexes

**Thèse dirigée par**
Raphaëlle ROY et Aurélie CLODIC

**Composition du jury**
M. Fabien LOTTE, Président, Université de Bordeaux
Mme Adriana TAPUS, Rapporteure, ENSTA Paris
M. Sylvain CHEVALLIER, Rapporteur, Université Paris-Saclay
Mme Wafa JOHAL, Examinatrice, University of Melbourne
Mme Raphaëlle ROY, Directrice de thèse, ISAE-SUPAERO
Mme Aurélie CLODIC, Co-directrice de thèse, LAAS-CNRS

# Abstract

Human–robot collaboration (HRC) is increasingly important in industrial, medical, and everyday settings. For robots to collaborate effectively, they must not only perceive the environment but also understand their human partner's mental states. This need has driven interest in processes that combine subjective, behavioral, and physiological data to estimate cognitive or affective states, a practice known as Mental State Monitoring (MSM), Physiological Computing, or Brain–Computer Interfaces (BCIs). This PhD explores the foundations, challenges, and possible contributions of MSM for HRC.

The work begins by reviewing current developments and challenges in HRC, with a focus on user evaluation methods. It highlights the value of physiological measures and the importance of assessing mental states during interaction, particularly cognitive effort, stress, and surprise. These are key indicators which allow to adapt robotic behavior and improve collaboration. It then provides an overview of the physiological signals most frequently used in MSM: brain, cardiac and electrodermal activity. It presents the origins of these signals, the sensors used to collect them and which features are typically extracted. This focus on standard physiological analysis is not limited to HRC applications and aims at offering a structured reference. MSM is subsequently formalized and broken down in key components. A distinction is made between offline, online, and pseudo-online analysis modes, each associated with specific trade-offs. Attention is given to both common and signal-specific noise sources, and how these can be mitigated through preprocessing techniques. Biomarkers of cognitive effort, stress, and surprise are reviewed, followed by a discussion of classification methods ranging from threshold-based methods to advanced machine learning and deep learning. The possible uses of MSM to guide robot adaptation is also presented at the levels of action, task planning and communication.

This foundation leads to a series of contributions. The first set aims at characterizing practical challenges such as intra-session variability and artifacts introduced by robot movement. A complete pipelines is introduced

for cognitive effort classification comparing three kind of physiological signals. These tools aim at bringing MSM closer to real-world application. A strong emphasis is then placed on robust experimental design. Methodological biases in MSM during HRC are identified and strategies are proposed to enhance its reliability. Three studies are presented: one validates several tasks inspired from neuropsychology for inducing cognitive effort through robot behavior alone using video-clip, the second uses these tasks to build a public dataset combining subjective, behavioral, and physiological data over two sessions per participant, and the third extends the paradigm to collaboration with a real robot. These efforts provide an empirical foundation for reproducible MSM research in HRC. Finally, two scenarios shift the focus toward surprise monitoring in HRC. One demonstrates perspectives for online monitoring using a public dataset while the other explores how humans detect errors in a complex collaborative task where the robot occasionally makes incorrect decisions. These experiments illustrate how MSM can be deployed for a second mental state, thus offering a more contrasted perspective on its application to HRC.

In summary, this PhD work builds a comprehensive framework for integrating mental state monitoring into human-robot collaboration. By addressing theoretical, technical, and methodological aspects, it lays the groundwork for future adaptive robotic systems that are sensitive to human cognition and capable of improving joint performance across domains. A strong emphasis is placed on open science, and the materials, datasets, and tools developed during this work are or will be made publicly available to support future research.

# Acknowledgements

A PhD is one of the few ways to contribute to the society as a scholar. It brings you to a strange world where you must cite your sources, give proofs of what you believe and where failure is useful. A world where your main task is to ask and answer questions, to the best of your abilities.

Thank you Raphaëlle and Aurélie for giving me the opportunity to get into this world. Merging physiological computing and robotics into one single project created a fantastic playground that I was glad to explore with you.

At the other end of the process, I would also like to deeply thank the jury. Both the reviewers, Adriana Tapus and Sylvain Chevallier, as well as the examiners, Wafa Johal and Fabien Lotte, are researchers whose work I admire. It is my honour that you took the time to give me your opinion on these three years.

In between, my life would not have been the same if I did not met Caroline, Maëlis and Marcel. Thank you for being wonderful colleagues, friends and human beings. Even if one of you is a bit less amazing than the two others.

To be honest, I should thank every past and present member of the CNE, including my interns, for offering me the best working environment one could possibly dream of. Naming all of you would make this paragraph way too long but you all are, or will be, great researchers and I am happy to pass half of my waking time in your company.

I do not forget the RIS team either, with a special mention for Guillaume, Bastien and Adrien. And a thought for Simon. It has been really inspiring to work with you and I hope we will keep in touch.

Thank you to all the enthusiastic people I met along the way, via ISAE-SUPAERO, LAAS-CNRS, Défi Clé, EDSYS, CORTICO, summer schools and all these great institutions or events which show that the true meaning of intelligence is to be shared.

Finally, thanks to my friends and family for filling the other half of my waking time with fun and joy. I love you all.

# Contents


**1 Introduction** **9**

**I State-of-the-art** **12**

**2 Decoding Humans for HRC** **13**
2.1 Collaborating with Robots . . . . . 13
2.2 Measurement Techniques . . . . . 15
2.3 Mental States of Interest . . . . . 17
2.3.1 Cognitive effort . . . . . 18
2.3.2 Stress and arousal . . . . . 19
2.3.3 Surprise and error detection . . . . . 20

**3 Electrophysiological signals** **21**
3.1 Brain activity . . . . . 21
3.1.1 Sensors . . . . . 22
3.1.2 Features . . . . . 25
3.2 Cardiac activity . . . . . 28
3.2.1 Sensors . . . . . 28
3.2.2 Features . . . . . 32
3.3 Electrodermal activity . . . . . 33
3.3.1 Sensors . . . . . 33
3.3.2 Features . . . . . 34

**4 Mental State Monitoring** **37**
4.1 Offline and online analyses . . . . . 38
4.2 Signal preprocessing . . . . . 39
4.3 Relevant features . . . . . 41
4.3.1 Cognitive effort . . . . . 42
4.3.2 Stress . . . . . 44
4.3.3 Error detection . . . . . 46

4.4 Classification . . . 47
4.5 Adaptation . . . 49
4.6 Scope of Contribution . . . 49

**II Contributions** **51**

**5 Make it work** **52**
5.1 Motivation . . . 52
5.2 Overview . . . 53
5.3 Methods . . . 54
5.3.1 Robot . . . 54
5.3.2 Physiological Sensors . . . 54
5.3.3 Software . . . 55
5.3.4 Tasks . . . 55
5.4 User Study . . . 56
5.4.1 Participants . . . 56
5.4.2 Protocol . . . 56
5.4.3 Data preprocessing and feature extraction . . . 58
5.5 Statistical Results . . . 60
5.5.1 Sequence length effect . . . 61
5.5.2 Session effect . . . 61
5.5.3 Interaction effect . . . 62
5.6 Classifying Cognitive Effort . . . 62
5.6.1 Pipelines . . . 62
5.6.2 Classification Results . . . 64
5.7 Robot noise . . . 67
5.7.1 Identification . . . 67
5.7.2 Preprocessing . . . 68
5.7.3 Classification . . . 68
5.7.4 Preprocessing Effect . . . 68
5.8 Discussion . . . 69

**6 Make it meaningful** **73**
6.1 Motivation . . . 73
6.2 Related Work . . . 75
6.3 RITHMB: Validation of an Open Access Protocol . . . 76
6.3.1 Overview . . . 76
6.3.2 Stimuli . . . 77
6.3.3 Tasks . . . 79

6.3.4 User studies . . . 83
6.3.5 Statistical Results . . . 85
6.3.6 Main outcomes . . . 91
6.4 RITHMD: a Physiological Database . . . 91
6.4.1 Overview . . . 91
6.4.2 Method . . . 92
6.4.3 User Study . . . 94
6.4.4 Statistical Results . . . 96
6.4.5 Classification Results . . . 105
6.4.6 Main outcomes . . . 108
6.5 RITHME: Comparison with a Physical Robot . . . 109
6.5.1 Overview . . . 109
6.5.2 Method . . . 109
6.5.3 User Study . . . 111
6.5.4 Statistical Results . . . 113
6.5.5 Classification Results . . . 117
6.5.6 Main outcomes . . . 118
6.6 Discussion . . . 119
6.7 Availability of data and materials . . . 121

**7 Contrasting with an other mental state 122**
7.1 Motivation . . . 122
7.2 Classifying Error . . . 123
7.2.1 Overview . . . 123
7.2.2 Data segmentation . . . 124
7.2.3 Pipeline . . . 124
7.2.4 Sample selection . . . 125
7.2.5 Pseudo-online detection . . . 126
7.3 DirectError . . . 127
7.3.1 Overview . . . 127
7.3.2 Director Tasks . . . 127
7.3.3 Methods . . . 129
7.3.4 User Study . . . 135
7.4 Statistical Results . . . 139
7.5 Classification Results . . . 140
7.6 Discussion . . . 140
7.7 Availability of data and materials . . . 142

**III General Discussion** **143**

**8 Discussion** **144**
8.1 A PhD about experimental design . . . 144
8.2 Comparing experimental designs . . . 145
8.2.1 Dual-task paradigms . . . 146
8.2.2 Neuropsychology-inspired HRC scenario . . . 146
8.2.3 Using a baseline task . . . 147
8.2.4 Comparative Insights . . . 147
8.3 Statistical insights on subjective, behavioral, and physiological metrics . . . 148
8.3.1 Subjective metrics . . . 150
8.3.2 Behavioral metrics . . . 150
8.3.3 Physiological metrics . . . 151
8.4 MSM insights from physiological data . . . 153
8.4.1 Brain activity . . . 154
8.4.2 Cardiac activity . . . 155
8.4.3 Electrodermal activity . . . 157
8.4.4 Summary about physiological activity . . . 158
8.5 Cognitive effort vs. error detection . . . 159
8.6 Perspectives . . . 160
Peer-Reviewed Contributions . . . 162
AI Statement . . . 162

# Chapter 1

# Introduction

Human–robot collaboration (HRC) is a challenge of increasing importance with applications in industry, medicine, and daily life settings. Unlike traditional automation, HRC is characterized by close physical and cognitive interactions, where the robot is not simply a tool but an adaptive partner. To ensure efficiency, safety, and user acceptance, robots must therefore be able to process not only the external environmental demands, but also the human partner's internal states.

Among these internal states, cognitive effort, stress, and surprise are particularly studied in the field. Cognitive effort reflects the mental resources invested in a task, both influencing performance and fatigue. Stress impacts decision-making, trust, and long-term well-being. Surprise, on the other hand, is a key marker of how humans detect and respond to unexpected robot behaviors. Accurately estimating these states in real time could enable robots to adjust their pace, redistribute tasks, or recover from errors in ways that enhance collaboration and user experience.

Physiological measures such as electroencephalography (EEG), electrocardiography (ECG) and electrodermal activity (EDA) offer a promising window into these mental states. Unlike self-report questionnaires, they provide objective, continuous and direct indicators of human experience. Their integration into HRC has the potential to transform robots into more responsive and socially attuned partners. However, while physiological monitoring has been widely explored in psychology, neuroscience, and human–computer interaction, to this day its application in HRC remains limited and fragmented.

The state of the art reveals several obstacles preventing the systematic use of physiological monitoring in HRC. First, the literature is sparse, with only a handful of studies targeting mental state monitoring in collaborative contexts. Second, the absence of standardized experimental paradigms makes results difficult to compare and replicate. Most researchers build their own

setups, often reinventing experimental tasks and feature extraction pipelines. Finally, physiological data are inherently variable, influenced by individual differences, sensor limitations, and environmental constraints. These challenges hinder the development of robust methods and prevent cumulative progress in the field.

This PhD thesis addresses these gaps by exploring how physiological measures can be systematically integrated into HRC research and by contributing to the establishment of common methodological ground. Specifically, it addresses the following research questions :

- What kind of HRC questions do physiological measurements help answer?
- To what extent do HRC constraints challenge the robustness of conventional MSM pipelines?
- Which experimental design allows to take the most insight from both worlds ?
- How can open research practices contribute to the long-term development of a growing research field?

To address these questions, the first three chapters establish the conceptual and methodological foundation. Chapter 2, Decoding Humans for HRC, examines why and how the human partner is evaluated within collaborative robotics, with particular emphasis on the role of mental states. Chapter 3, Electrophysiological Signals, introduces the physiological measures employed in this work: brain, cardiac, and electrodermal activity. For each signal, it presents the underlying physiology, the sensors available, and common methods of analysis. Chapter 4, Mental State Monitoring, builds upon the previous chapters to discuss specific processes involved in this field. By examining both the potential and the limitations of MSM for HRC, it sets the boundaries within which the contributions of this PhD are positioned.

The next three chapters present empirical contributions. Chapter 5, Make it Work, investigates the feasibility of MSM in HRC by studying electrophysiological markers of cognitive effort under real-world constraints such as inter-individual variability, wearable sensor limitations, and artifacts introduced by robot. Chapter 6, Make it Meaningful, turns to experimental design, introducing neuropsychology-inspired paradigms that enable the study of cognitive effort in reproducible HRC scenarios. Chapter 7, Contrasting with Another Mental State, extends the focus beyond cognitive effort to explore the detection of surprise in HRC, thereby offering a complementary

perspective on MSM. Finally, the General Discussion synthesizes the insights gathered across the PhD, proposing concrete methodological guidelines and providing open datasets to support the design of future research protocols in this growing field.

# Part I

# State-of-the-art

# Chapter 2

# Decoding Humans for HRC

## 2.1 Collaborating with Robots

In the future, robots hold the promise of offering assistance in numerous aspects of our daily lives and fulfilling key societal functions. Their potential as social agents has already been explored in diverse ecological settings, taking on roles such as restaurant waiters [1], shopkeepers [2], or shopping mall guides [3] (see Figure 2.1). These real-world deployments demonstrate both the promise of human-robot interaction (HRI) and the diversity of subtasks that robots will have to perform to collaborate smoothly with humans and gain acceptance within society.

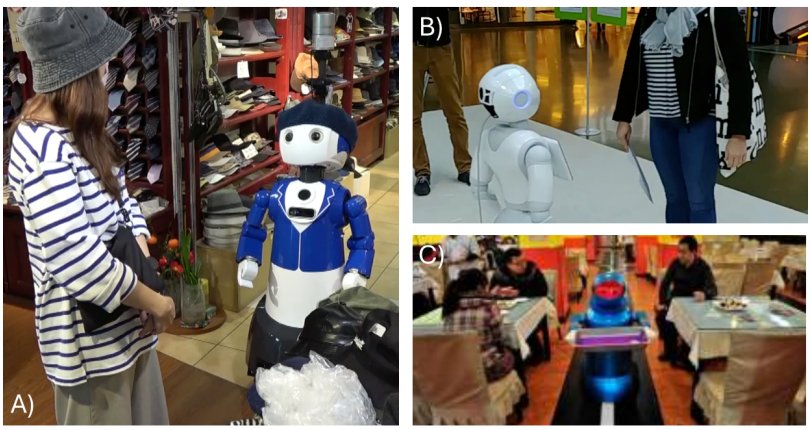


Figure 2.1: Examples of real-world user studies involving collaborative robots: **A)** a robot assisting customers in a hat shop [2], **B)** a social robot deployed in a shopping mall [3], and **C)** a service robot bringing food in a restaurant [1].

As the field of HRI evolves, it increasingly demands a deeper integration of human factors in both the design and evaluation of robotic systems. Standardized and replicable methodological approaches are critical not only for ensuring robustness and reproducibility but also for motivating active debates within the research community [4, 5]. However, capturing the richness of real-world interaction in a single evaluation remains a challenge. Ideally, user studies that aim to characterize the use of a robot in everyday life would involve a broad spectrum of participants across multiple sessions to control for individual differences and session-related variability. Given the significant resources required for such studies, research often begins with more targeted investigations in controlled environments. Yet, even within these simplified scenario, clearly articulating research questions and limitations is crucial to ensure transparency and reliability.

Focusing on human-robot collaboration (HRC) introduces additional cognitive and behavioral constraints. The joint action paradigm [6] emphasizes that successful collaboration depends on mechanisms such as joint attention, shared task representation, and the understanding of intentional actions. Robots do not need to imitate human behavior to collaborate effectively; rather, they must be capable of coordinating in ways that allow them to understand and be understood by humans [7]. This perspective calls for evaluation metrics that go beyond end-goal performance, encompassing additional dimensions of interaction that may otherwise be overlooked. For instance a complementary metric could be fluency, defined as a well-synchronized meshing of agents actions which allow to couple precise and efficient action timing with appropriate and dynamic adaptations of both agents without need for verbal communication [8].

In addition, as collaboration increasingly depends on mutual understanding, evaluating both humans and robots individually becomes a central challenge. Differences between human expectations and robotic behavior can undermine trust and fluency. While the robot's internal state is generally accessible, there is growing interest in leveraging psychological metrics to infer human mental states such as stress, attention, cognitive effort, or surprise [9, 10]. Understanding and adapting to these variations could allow robots to respond in ways that are not only efficient but also psychologically and socially attuned [11].

The integration of collaborative robots, commonly known as cobots, into industrial environments represents a major technological shift, enabling humans and robots to work side by side in shared workspaces. This context provides a fertile ground for research, where key questions related to human-robot physical collaboration are actively explored. Numerous studies aimed to optimize human-robot collaboration by investigating operator safety, in-

teraction fluency, and cognitive ergonomics within these settings [10, 12, 13]. In parallel, a growing body of work has focused on operator well-being, addressing factors such as trust, mental workload, and anxiety when interacting with cobots [14, 15]. Despite these advancements, much of the research has been conducted using robotic arms or other non-anthropomorphic platforms. As a result, important questions remain regarding the generalization of these findings to humanoid robots, which may elicit stronger social expectations and more nuanced behavioral responses from users [16, 17].

At the other end of the spectrum, social robots hold significant promise in domains such as education [18, 19] and healthcare [20, 21], where the focus shifts from task efficiency to developing empathy and communication. These robots primarily rely on verbal interaction with a humanoid design aimed at supporting embodied interaction. This allows for multimodal communication through facial expressions, gestures, gaze, and other social cues that enhance user engagement and interpersonal connection [22, 23]. In these contexts, understanding the human mental state is often approached through emotion recognition, based on the analysis of facial features, vocal tone, posture, and increasingly, physiological signals [24]. Yet, cognitive states should not be overlooked as they are of great interest to better understand both failures and performances in cooperative settings.

In this broad context, the present work focuses on scenarios that involve a humanoid robot, inducing high social expectations through its embodiment, exploiting the physical possibilities offered by its presence. By exploring how cognitive states can be evaluated during interaction with robotic systems, we aim to contribute to the design of collaborative robots that adapt dynamically to human needs. Grounded in concepts of fluency, adaptation, and human-centered design, this research ultimately supports the development of robotic partners that are not only functional but also natural, respectful, and effective in shared environments.

## 2.2 Measurement Techniques

Common measurement techniques in the HRC community include self-assessment, interviews, behavioral measures, task performance metrics and physiological markers [25].

Among them, subjective methods such as questionnaires and interviews remain the most popular for evaluating human mental states and assessing the quality of interaction. Standardized instruments like the Godspeed Questionnaire Series [26, 27], the Negative Attitudes Toward Robots Scale (NARS) [28] or the Human-Robot Interaction Evaluation Scale (HRIES) [29]

are widely used to evaluate users' impressions of robots such as anthropomorphism, likability, and discomfort. When it comes to assess overall user state, more generic questionnaires are often used. These include the NASA Task Load Index (Nasa TLX), which quantifies perceived cognitive workload [30], and the Positive and Negative Affect Schedule (PANAS), used to evaluate emotional well-being [31, 32]. While less frequently employed, qualitative approaches such as interviews and focus groups can also contribute to gather insights into the user's experience [33]. However, these subjective assessments are most often collected before and after task execution, or during scheduled breaks. This timing results in task interruptions and limits the temporal resolution of the data, making it difficult to monitor real-time changes or support adaptive behavior during the interaction. Consequently, these methods provide only a session-level understanding of the collaboration, highlighting the need for complementary, continuous approaches to better capture the dynamics of human-robot interaction. Indeed, to offer deeper insights into the interaction dynamics, the quality of interaction should also be evaluated regarding each task and each action [34, 35]. In addition, conducting this evaluation solely from the robot's perspective, without external assistance, would ensure its applicability across a wider range of scenarios.

To address such issues, objective metrics, particularly behavioral and task performance metrics, are a common complement to subjective evaluations in HRC studies. Among behavioral metrics, gaze or head tracking provides insights into users' attention, focus, and engagement [36, 37]. It has been used to evaluate whether robot behaviors are noticeable and socially engaging, for instance during interaction with a shopkeeper [38] or a museum guide [39]. Gesture recognition is another key behavioral indicator which helps understand how naturally and efficiently users interact with robots. Systems utilizing RGB-D sensors or motion tracking technologies allow robots to identify, track and classify human gestures [40], facilitating adaptive communication in domains like elderly care [41]. Complementarily, task performance metrics quantify the effectiveness of collaboration. Task completion rate, execution time, and accuracy are widely used to measure whether shared goals are achieved efficiently and correctly, with a majority of studies demonstrating their relevance in industrial applications [10, 42]. Additionally, idle time - representing moments when one or both agents are not actively contributing — offers insights into coordination fluency and potential bottlenecks [8]. Together, these behavioral and performance indicators enable a more nuanced, real-time assessment of human-robot collaboration, supporting the development of adaptive systems capable of improving dynamically the quality of interaction. Yet, they may offer limited insights into human mental states when compared to subjective metrics.

To bridge the gap between observable behaviors and the underlying human state, physiological metrics have emerged as valuable tools in HRC studies. These measures offer direct insights into users' cognitive and emotional states, often beyond what can be captured through behavior or performance alone. For instance, electroencephalography (EEG) can provide real-time data on cognitive workload, attention, and error detection by monitoring brain activity[43]. While providing less complex signals, cardiac, muscular, and electrodermal activity can also reflect in their variations different levels of emotional arousal or cognitive load [44]. In addition, changes in pupil diameter has also proven useful in assessing moment-to-moment shifts in attention and affect [45, 46]. While the interpretation of physiological signals requires careful consideration of individual and contextual variability, their integration into HRC research enables continuous monitoring of users' mental states through objective measurement. This not only complements behavioral and performance measures but also paves the way for more responsive robotic systems.

Thus, this work plans to collect subjective, behavioral and physiological measurement during human-robot interaction scenario to better characterize the added value of physiological signals and the new paradigms that they bring. It will particularly focus on the evaluation of human cognitive state during these interactions.

## 2.3 Mental States of Interest

Advancements in neuroscience and the availability of portable physiological monitoring devices have significantly expanded the applicability of physiological metrics across a range of domains, including HRC. From a human factor perspective, the emerging field of neuroergonomics [47, 48] offers a powerful framework for evaluating the mental states of human operators engaged in sustained interactions with complex systems through the use of neuroscientific paradigms. More recently, the integration of machine learning into these approaches has given rise to physiological computing, a dynamic way to interpret biosignals to adapt system behavior [49]. These developments enable the creation of a retroactive physiological loop wherein continuous physiological feedback allows systems to adapt their responses in real time.

Despite the promising potential of these technologies for HRC, current physiological sensors and decoding methods are still limited in their ability to fully interpret the many dimensions that constitute a user's mental state. Therefore, building an effective retroactive loop begins with a careful selection and investigation, both offline and online, of a specific mental

state. While the detailed methodology for this process will be outlined in another chapter (see Section 4), the following section provides an overview of the mental states most frequently identified as critical for enhancing user experience, task performance, and safety in HRC settings.

### 2.3.1 Cognitive effort

One of the most extensively studied is cognitive effort, also referred to as mental engagement or mental workload [50]. It is related to the tendency of human performance to decline within a task particularly challenging or, in contrast, overly simple, following a U-shape curve [51]. Detecting fluctuations in cognitive effort is essential for identifying instances of mind wandering, where individuals disengage from the interaction, as well as periods of cognitive overload, characterized by a lack of fluency in the interaction or the need for a cognitive break [52].




Figure 2.2: Example of a worker collaborating with a robot on a construction site, where the worker's cognitive effort is monitored through physiological measurements [53].

For instance, EEG has been effectively applied to assess workload during physical HRC scenarios, allowing robotic system to adapt its assistance based on fluctuations in cognitive demand [54]. It has also been used to adjust the pace of a rover base on worker cognitive effort [53] (See Figure 2.2). Multi-signal approaches have also shown promise, with wearable sensors being used to detect workload variations during complex assembling

tasks involving robotic arms [55, 56, 57, 58]. Altogether, these approaches underscore how monitoring cognitive effort can allow robots to dynamically adjust their behavior to maintain an optimal level of engagement and support users in completing tasks effectively [10, 15].

### 2.3.2 Stress and arousal

Stress and arousal represent another essential dimension which offer insights into human safety when sharing a workspace with robots. This can cover a wide range of parameters related to the prevention of any psychological or physical discomfort in HRC [59].

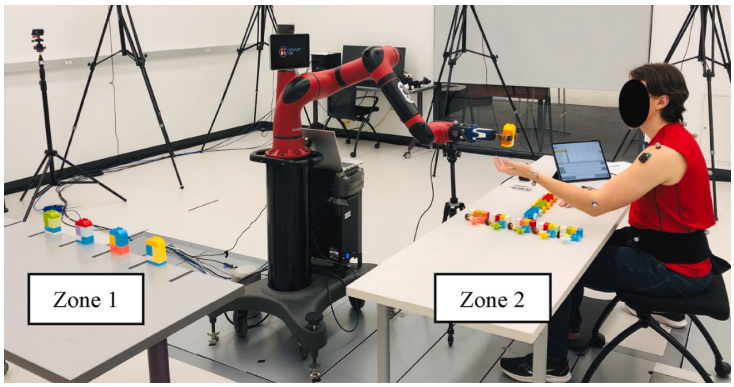


Figure 2.3: Example of a participant engaged in a collaborative assembly task with a robot in a shared workspace environment [60].

For instance, tactile interaction with humanoid robots has been linked to increased physiological arousal, with faster electrodermal activity suggesting heightened engagement and alertness [61]. Even with no contact, monitoring physiological activity has enabled to evaluate stress level in a shared workspace and adjust collaboration accordingly [60] (see Figure 2.3). It has also led to the evaluation of the impact of robot interaction style on user's comfort [62]. Additionally, robot behavioral features such as increased movement speed, dominant movements, unpredictability, or inadaptability of the robot have been shown to decrease trust perceptions and increase stress [14, 63]. These examples highlight how stress detection, when integrated into adaptive control strategies, can significantly improve both user experience and system performance, further reinforcing the importance of mental state monitoring in the broader landscape of cognitive adaptation in HRC.

### 2.3.3 Surprise and error detection

Surprise and error detection play a crucial role in HRC scenarios, as in other dynamic and unpredictable human-machine interactions [64]. Unexpected events, which can be qualified as novel or erroneous depending on the context [65], may disrupt the fluency of collaboration and challenge users' beliefs about the robot's capabilities or intentions. By detecting instances of surprise, systems can adapt their behavior to address discrepancies and align with human beliefs during the ongoing process [66].




Figure 2.4: Example of a participant engaged in a cueing task with a collaborative robot, in which the robot deliberately introduces errors to investigate human physiological response [67].

For instance, when a user detects a robot's mistake, a distinct neural response can be recorded to trigger immediate corrective actions (see Section 4.3.3). Such responses have been successfully used to create closed-loop systems where the robot adjusts its behavior in real time based on human recognition of errors, improving both safety and efficiency in collaborative tasks [67, 68] (see Figure 2.4). Other approaches combine brain and muscle signals to detect gesture-related anomalies, enabling supervisory systems to recognize user uncertainty or incorrect execution and intervene proactively [69]. It has also been used as a human implicit feedback to train robots to select the most suitable behaviour through reinforcement learning techniques [70]. These developments suggest that integrating physiological feedback into HRC systems not only enhances error correction mechanisms but also supports smoother, more intuitive interactions by aligning robotic behavior with human expectations.

# Chapter 3

# Electrophysiological signals

Electrophysiological signals offer a direct and dynamic window into the functioning of the human nervous system by capturing the subtle electrical fluctuations generated by biological activity. Emerging from diverse physiological processes, these signals convey rich information about both bodily states and variations in mental activity.

This chapter presents the primary electrophysiological signals relevant for mental state monitoring (MSM) in Human-Robot Collaboration (HRC). It outlines the sensors employed to acquire them as well as the characteristic features typically extracted for analysis, thereby laying the groundwork for the methodological approaches developed in the subsequent sections.

To facilitate navigation across the different signals, sensors, and features described, Table 3.1 provides a comparative overview of the sensors and their key characteristics.

| Acronym | Recorded Activity | Localisation | Sensors |
|---|---|---|---|
| EEG | Brain | Head | 1 to 256 |
| ECG | Cardiac | Torso, Arm, Leg | 2 to 12 |
| PPG | Cardiac | Finger, Ear, Wrist | 1 |
| EDA | Electrodermal | Finger, Palm | 2 |

Table 3.1: Summary of discussed electrophysiological sensors.

## 3.1 Brain activity

Monitoring brain activity involves capturing the coordinated dynamics of large populations of neurons. This activity reflects fundamental processes such as perception, attention, and memory, and its fluctuations can reveal

variations in human mental state. As such, brain monitoring provides a direct and dynamic window into the core of human perception.

### 3.1.1 Sensors

Among the available physiological signals, brain activity stands as the richest, which also makes it the more complex to record, process and interpret. The most popular sensors include electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG) (see Figure 3.1). EEG is the most widely used for wearable applications due to its high temporal resolution and relatively low cost. It works by detecting electrical signals from the scalp generated by neuronal activity, typically using surface electrodes that can be integrated into lightweight, portable devices such as headbands or caps [71]. Despite its low spatial resolution and susceptibility to noise from motion or muscle activity, EEG remains ideal for brain monitoring in mobile context. In contrast, fNIRS, which measures changes in blood oxygenation via infrared light, offers better spatial resolution than EEG and is portable, but it suffers from lower temporal resolution and high sensitivity to light conditions. fMRI, known for its excellent spatial resolution and whole-brain coverage, relies on detecting blood flow changes but is unsuitable for wearable use due to its high cost, immobility, and poor temporal resolution. MEG provides both high spatial and temporal resolution by measuring magnetic fields from neural activity, yet its dependence on cryogenic equipment and shielded rooms makes it impractical for portable use. Overall, EEG stands as the only non-invasive solution to collect brain activity with a sufficient temporal resolution during mobile experiments [72].

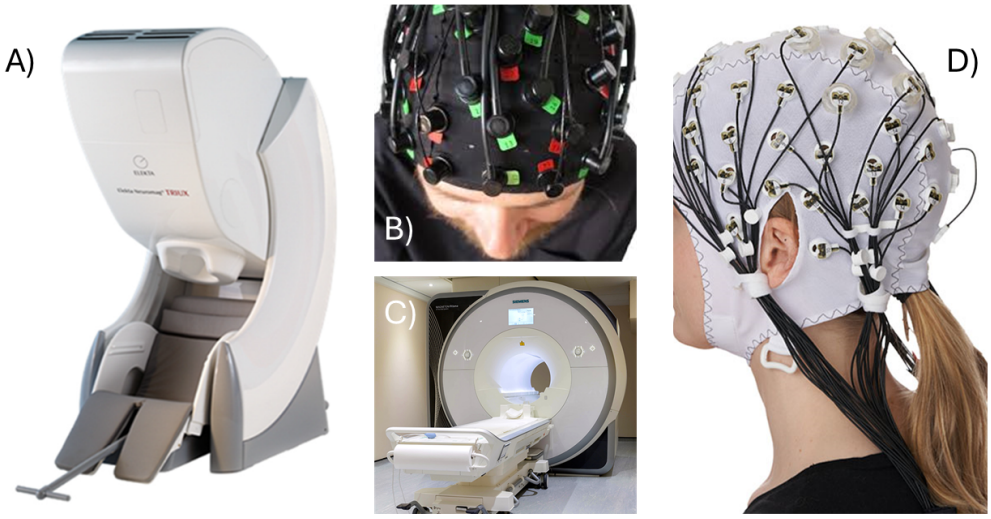


Figure 3.1: Illustrations of major brain sensors. **A)** Magnetoencephalography (MEG), capturing magnetic fields generated by neural activity. **B)** Functional near-infrared spectroscopy (fNIRS), measuring changes in blood oxygenation at the cortical surface. **C)** Functional magnetic resonance imaging (fMRI), providing whole-brain hemodynamic activity with high spatial resolution. **D)** Electroencephalography (EEG), recording scalp electrical activity with high temporal precision.

Because EEG systems vary widely in design and capabilities, selecting the right system depends heavily on the intended application, whether for clinical diagnostics, laboratory research, or mobile, real-world monitoring. To make informed decisions, it is essential to understand the key parameters that influence EEG performance, usability, and data quality.

One of the most fundamental choices is the number of channels, as this directly determines the spatial resolution of the recorded brain activity. High-density systems, typically comprising 64 to 256 channels, offer superior spatial coverage and are well-suited for clinical diagnostics or fundamental neuroscience research where detailed cortical mapping is essential. In contrast, low-density systems with 1 to 32 channels are commonly used in brain-computer interface (BCI) applications and mobile contexts due to their reduced setup complexity, improved portability, and lower computational requirements.

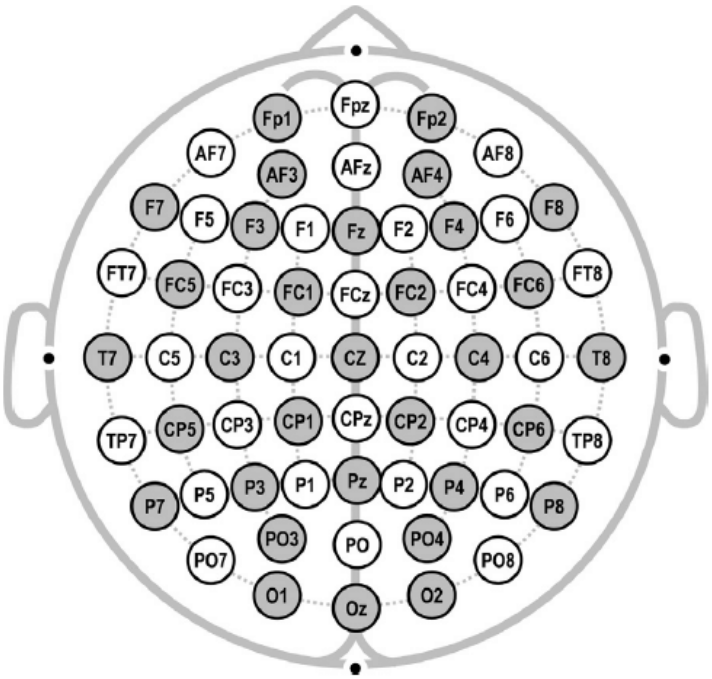


Figure 3.2: Illustration of the international 10–20 system for EEG electrode placement with 64 channels. Grey circles indicate the subset corresponding to a 32-channel configuration [73]. Electrode labels combine a letter, denoting the underlying brain region (e.g., F for frontal, P for parietal), and a number, denoting lateral position. Odd numbers correspond to the left hemisphere and even numbers to the right hemisphere.

Beyond quantity, electrode placement strategy also requires thoughtful design. Most systems adhere to the international 10/20 or 10/10 placement standards (see Figure 3.2), which ensure methodological consistency and comparability across studies. However, when investigating specific brain regions, researchers may opt to increase electrode density locally while reducing it elsewhere to optimize spatial focus. While custom arrangements can enhance signal acquisition from targeted areas, diverging from standard layouts must be done cautiously to maintain interpretability and comparability[74].

Another key parameter is the electrode type. Wet electrodes are widely regarded as the gold standard for signal fidelity, offering low impedance and high-quality data. However, they require time-consuming preparation, including the application and removal of conductive gels, making them less practical for mobile or everyday use. In contrast, dry and hybrid electrodes offer faster setup, which is ideal for wearable systems, though they can suffer from inconsistent contact and lower signal quality [75, 76].

The sampling rate also plays an essential role, as it determines the temporal resolution of the EEG signal. High sampling rates, usually above 1000 Hz, can be necessary for capturing rapid neural dynamics and some event-related potentials but generate larger datasets and demand greater processing capabilities. For standard applications, a rate of 250–500 Hz often provides a suitable balance between temporal resolution and system efficiency[77].

Lastly, portability has become increasingly important as EEG moves beyond the laboratory. Modern wearable EEG systems prioritize minimal weight, wireless communication, and battery lifetime, to facilitate naturalistic movement and reduce encumbrance during recordings [78]. Yet, maintaining high signal quality in such context remains a technical challenge due to unstable connection, increased motion artifacts and external interferences.

In summary, choosing an EEG system involves balancing spatial and temporal resolution, signal quality, portability, and user comfort. The optimal configuration depends on the specific goals and constraints of the intended application—whether it's high-resolution neural mapping in a clinical setting or lightweight, unobtrusive monitoring in everyday environments.

As this work focuses on field applications, a wearable wireless EEG system with a low number of (dry) electrodes will be used most of the time.

### 3.1.2 Features

EEG signal can be analyzed through a broad array of extracted features which fall mostly into three primary categories: temporal, frequency and spatial. Each of these categories provides a unique perspective into brain function and can support a variety of applications.

Temporal features, particularly event-related potentials (ERPs), play a crucial role in EEG analysis by capturing the brain's time-locked responses to specific sensory, cognitive, or motor events. Reference ERPs are derived by averaging EEG segments that are time-aligned to a stimulus, which enhances signal-to-noise ratio and isolates meaningful neural responses. Characteristic ERP components, peaks and troughs of EEG waves, such as the P300, N400, or N170, are defined by their polarity (N for negative, P for positive) and typical latency (in ms after stimulus). They are linked to distinct cognitive processes like attention for P300 [79] (see Figure 3.3), language comprehension for N400 [80] or facial recognition for N170 [81]. Key features extracted from ERPs include peak amplitude, latency, duration, and waveform morphology. While the overall ERP shape remains characteristic, these features can be impacted by human current mental state. For example, the amplitude of the P300 component, elicited by a rare stimuli, can vary according to task difficulty [82]. Hence, the millisecond precision provided by ERPs

allows researchers to track the dynamics of brain activity as it unfolds over time.

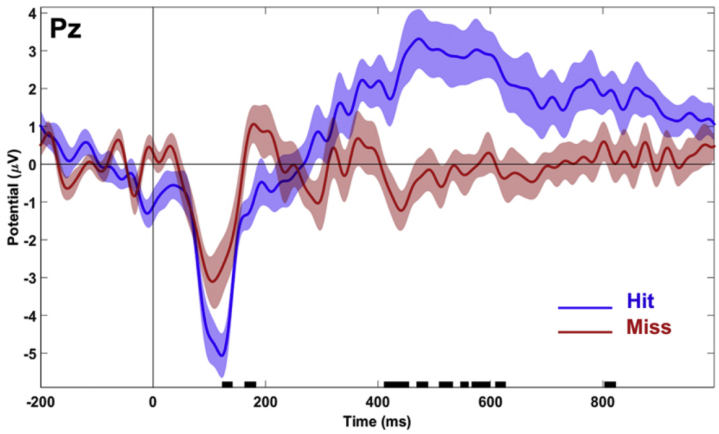


Figure 3.3: Typical P300 response elicited during an auditory oddball task [79]. Participants listen to a series of frequent (standard) and rare (target) sounds while performing the main task, pressing a button whenever the target sound is detected. Detecting the rare target stimulus reliably evokes a positive deflection around 300 ms (shown in blue).

Frequency domain features involve transforming the EEG signal into its constituent frequency components using techniques such as the Fourier Transform. One of the most widely used frequency features is the power spectral density (PSD), which quantifies the energy within specific frequency bands (see Figure 3.4). Common frequency bands of interest include delta $(0.5 - 4Hz)$, theta $(4 - 8$ Hz), alpha $(8 - 12$ Hz), beta $(12 - 30$ Hz), and gamma $(> 30$ Hz). These frequency bands have been associated with distinct cognitive processes and mental states. For example, delta oscillations have been linked to attention processes, theta oscillations to memory encoding and inhibition, while alpha oscillations have been associated with states of relaxation, inhibition, or sensory attenuation. Beta oscillations seem involved in motor control and sensorimotor integration, while gamma oscillations have been associated with higher cognitive functions such as information processing, conscious perception and memory retrieval [83]. Hence, PSD can be computed at several time points to measure the power of a certain frequency band of the signal and thus the evolution of a correlated mental states across time and/or conditions. This power metric can be used absolute, relative to the global spectrum or as a ratio, as with the engagement ratio $\frac{\beta}{\alpha+\theta}$ [84]. Other features such as sample entropy and approximate entropy evaluate the signal's complexity to offer a complementary view [85].

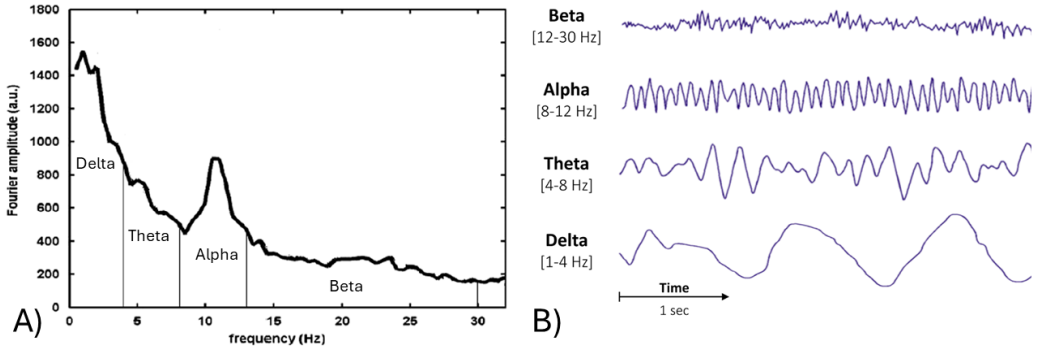


Figure 3.4: Canonical frequency bands commonly used in EEG analysis. (A) Example of a typical power spectrum of brain activity at rest, highlighting the dominant frequency ranges. (B) Illustration of the temporal waveforms corresponding to each frequency band when isolated.

Spatial features describe how neural activity is distributed across the scalp and how different brain regions functionally interact. Although EEG has limited spatial resolution compared to other imaging modalities, spatial information remains highly valuable for the characterization of brain processes. One of the most impactful spatial descriptors in EEG analysis is the covariance matrix, which captures the statistical relationships between signals recorded from multiple electrodes. When analyzed using Riemannian geometry, covariance-based approaches have led to significant advances in brain-computer interface performance [86]. On the other hand, functional connectivity measures assess the coordination and communication between different brain areas. These metrics include coherence, phase synchronization, the phase-slope index, and Granger causality, each quantifying different aspects of inter-regional rhythmic activity [87, 88].

It is important to recognize that the boundaries between temporal, frequency, and spatial features are not rigid, as many EEG features inherently combine information from multiple domains. For instance, region-of-interest (ROI) analysis involves selecting specific sets of electrodes that correspond to known anatomical or functional brain regions, such as frontal, parietal, or occipital lobes, and extracting their central tendency. Thus, any of the previously cited temporal or frequency feature can be processed at the ROI level instead of the electrode level, integrating spatial prior to the analysis. Similarly, ERP analysis can be extended to the frequency domain to uncover underlying oscillatory dynamics [89]. This kind of time-frequency analysis, though requiring heavy computational power, provides deeper insights into

brain activity.

Finally, the increasing integration of machine learning techniques in EEG analysis has given rise to data-driven features that reflect complex patterns learned from the data. Among the most prominent is Common Spatial Pattern (CSP), which identifies spatial information that maximize the variance between two conditions, effectively enhancing the signal-to-noise ratio for discriminative brain activity [90, 91]. Another noteworthy approach is the microstate analysis [92, 93], which applies clustering algorithms, commonly k-means, to identify recurring topographical patterns in EEG signals. These microstates are thought to reflect brief, fundamental units of spontaneous brain activity, whose temporal dynamics can reveal transitions between cognitive states. Although such features can significantly enhance brain signal decoding, their dependence on the training data and the variability of the learned patterns challenge traditional definitions of EEG features, pushing the boundary between feature extraction and modeling. For conceptual clarity and coherence, these methods will be further discussed in the chapter devoted to mental state monitoring in direct relationship with the targeted applications.

## 3.2 Cardiac activity

Monitoring cardiac activity focuses on capturing the rhythmic electrical and mechanical processes that drive the heartbeat, starting with contractions in the upper chambers and propagating through the lower chambers. Because cardiac function is closely regulated by the autonomic nervous system, variations in a person's mental state can directly influence heart activity, making it a valuable window into psychophysiological processes.

### 3.2.1 Sensors

Electrocardiography (ECG) remains the gold standard for monitoring cardiac activity, providing a direct measurement of the heart's electrical signals through electrodes placed on the skin, typically across the chest, arms, and legs. These signals capture the depolarization and repolarization of the heart muscle, reflecting each stage of the cardiac cycle. In clinical settings, ECG often integrate 12 channels to enable detailed, three-dimensional analysis of cardiac function and diagnose various heart conditions with high accuracy [94]. However, for applications such as human performance monitoring, the emphasis shifts from diagnostic precision to portability and ease of use. As a result, the number of electrodes is often reduced, with configurations lowering

down to 2 channels, as they may be sufficient for simple heart rate monitoring. Thus, depending on the application, ECG can require the placement of 2 to 12 adhesive electrodes [95] (see Figure 3.5). Some highly portable systems even replace the individual adhesive electrodes with conductive materials embedded in a band worn around the chest. These chest straps offers low-cost alternatives which may provide better motion artifact resistance, making it a practical solution for real-time monitoring in dynamic environments like sports.

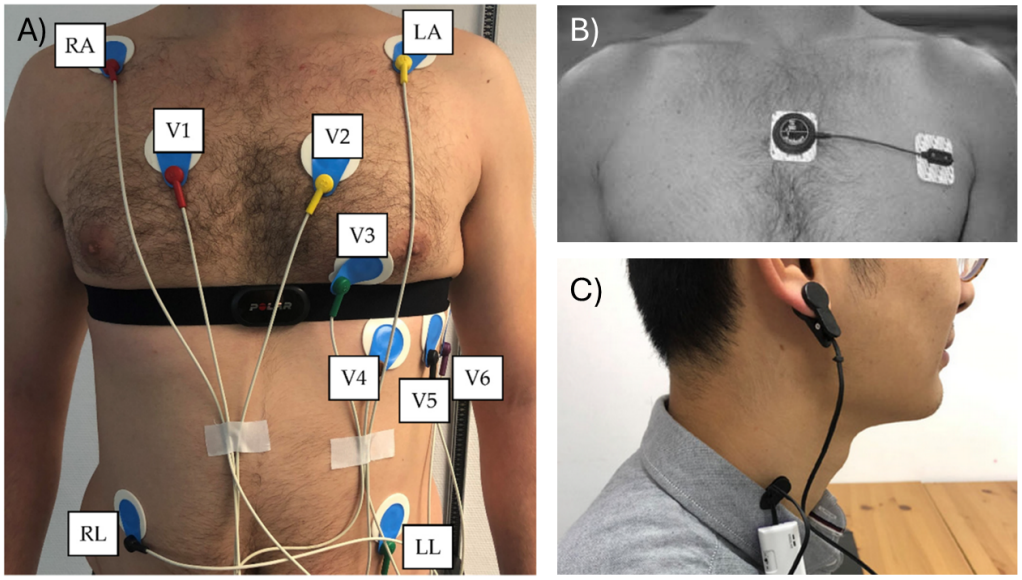


Figure 3.5: Illustrations of different cardiac sensing modalities. **A)** A participant wearing a clinical-grade 12-lead ECG system together with a Polar H10 chest strap [96]. **B)** A participant equipped with a lightweight 1-channel ActiHeart ECG device [97]. **C)** A participant using an ear-clip photoplethysmography (PPG) sensor [98].

As an alternative, photoplethysmography (PPG) is an optical method for monitoring cardiac activity by detecting blood volume changes in the microvascular tissue using a light source and photodetector. With each heartbeat, these volume changes affect the absorption or reflection of light, producing a waveform that reflects the cardiac pulse and allows for heart rate estimation. While PPG is less precise than ECG and more susceptible to motion artifacts and ambient light, it is highly adaptable and easy to wear [99]. The PPG sensor can be placed on various body sites such as wrist, fingertip, or earlobe, and newer techniques like remote PPG (rPPG) even allow

contactless measurement via standard cameras, though they raise privacy and ethical concerns [100, 101]. This comfort and convenience led the PPG to become a standard component of commercial wearables, such as smartwatches and smart rings, enabling daily, real-world monitoring for a growing number of people. Yet, most commercial devices limit research use due to proprietary algorithms and restricted access to raw data. Thus research-grade PPG sensors remain less user-friendly by comparison. Still, ongoing advances in sensor design and signal processing are improving PPG's reliability, reinforcing its role as a highly-wearable tool for non-clinical, continuous heart monitoring.

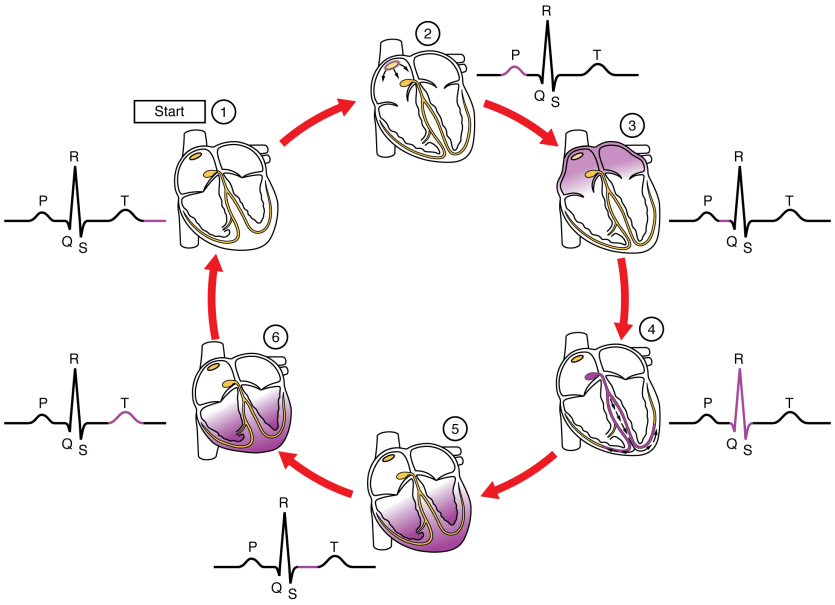


Figure 3.6: Illustration of the QRS complex and its correspondence to cardiac activity. The electrical waveform reflects the depolarization of the ventricles, mapped onto the mechanical contraction of the heart.

Hence, both ECG and PPG are sensitive to the same underlying phenomenon but they measure cardiac pulsation through fundamentally different mechanisms. ECG provides a high-resolution view of the heart's electrical activity, producing a detailed waveform that captures the sequential stages of the cardiac cycle. Its central feature, the QRS complex, includes two deflections (Q, S) and a peak (R) induced by ventricular depolarization (see Figure 3.6). This is typically preceded by the P wave, resulting from atrial depolarization, and followed by the T wave, induced by ventricular repolarization. In human-performance monitoring, analysis often centers on

detecting R peaks and calculating RR intervals (the time between successive R peaks), often normalized and renamed NN intervals, which serve as the foundation for extracting key cardiac features [102], as it will be detailed in the features subsection (see Section 3.2.2).

In contrast, PPG records peripheral blood volume changes, producing a characteristic, though less detailed, waveform. It begins with a systolic peak, representing the surge in blood volume following ventricular contraction, followed by a dicrotic notch and a diastolic peak before returning to baseline. The timing between successive systolic peaks, known as the PP interval, is used similarly to RR intervals to derive cardiac features [103]. Because PPG captures the peripheral effects of cardiac activity, it introduces a delay, known as pulse transit time (PTT), of several hundred milliseconds between the electrical activation observed in ECG and the resulting pulse detected by PPG (see Figure 3.7). While generally stable in healthy individuals at rest, this delay can fluctuate with physiological or environmental changes, leading to subtle discrepancies between PP and RR intervals [104, 105].

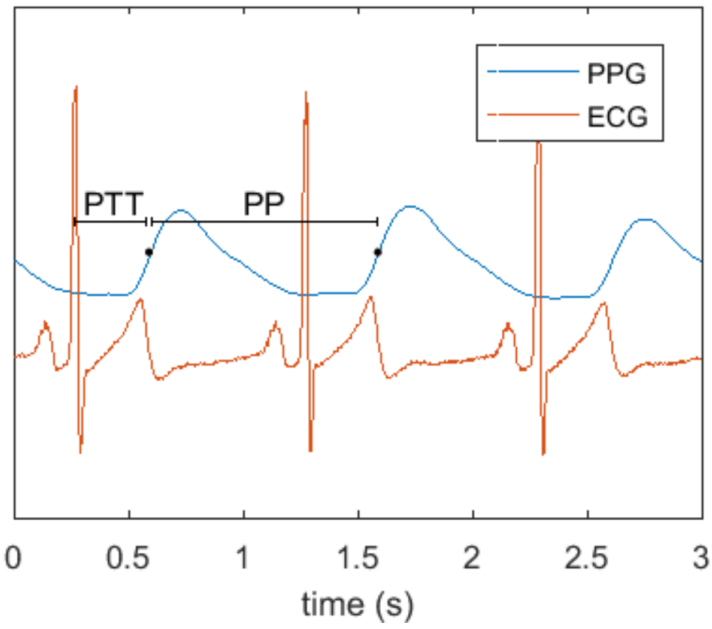


Figure 3.7: Illustration of the delay between R peak and P wave from Kubios [106].

Hence, while both modalities capture the same underlying cardiac events, their distinct measurement principles shape the precision, timing, and interpretation of the data they provide. Characterizing these differences is crucial

for evaluating the trade-offs between PPG's superior wearability and ECG's higher temporal precision. Accordingly, most experiments presented here employ both sensors to compare their strengths and limitations in diverse monitoring contexts.

### 3.2.2 Features

Both the RR-interval and the PP-interval allow us to derive similar metrics. These latter usually belong to one of the three following groups: time, frequency or non-linear [107].

In the time domain, features include heart rate (HR), directly computed from the inverse of the RR (or PP) intervals, and various heart rate variability (HRV) metrics. While HR correlates closely with overall cardiac demand, HRV provides a sensitive window into autonomic nervous system regulation. Among the most validated HRV measures are the standard deviation of NN intervals (SDNN), which reflect overall variability, and the root mean square of successive differences (RMSSD), which is particularly sensitive to short-term activity [108].

In the frequency domain, spectral analysis allows to study HRV within standard bands such as low frequency (LF, 0.04–0.15 Hz) and high frequency (HF, 0.15–0.4 Hz), with the LF/HF ratio commonly interpreted as a marker of sympathetic-parasympathetic balance [109].

Finally, non-linear features aim to capture the inherent complexity and irregularity of cardiac signals, going beyond assumptions of stationarity and linearity [110]. These include entropy-based measures that quantify unpredictability, Poincaré plot analysis that illustrates beat-to-beat dynamics and provides indices of short- and long-term variability, and fractal analyses that examine the self-similar structure of heart rate fluctuations over multiple time scales.

Collectively, these measures provide complementary perspectives on cardiac dynamics and their associations with both physical and psychological states. However, a key challenge concerns the minimal recording duration needed for reliable feature extraction. Most traditional HRV metrics are designed to estimate central tendencies over relatively long windows, typically requiring at least 5 minutes of continuous data to achieve robust results. It can go up to 10 minutes or more for low-frequency spectral components which need to capture slow oscillations accurately [111]. Nonetheless, recent research has demonstrated that certain time-domain metrics, particularly those sensitive to short-term variability such as RMSSD, can yield reliable estimates even in ultra-short recordings of 10–30 seconds [112, 107]. This opens promising avenues for real-time and ecological monitoring of cardiac

activity in dynamic settings, though caution remains necessary when interpreting spectral and nonlinear indices over such brief intervals.

## 3.3 Electrodermal activity

Electrodermal activity (EDA) refers to the dynamic changes in the skin's electrical conductance, primarily driven by fluctuations in sweat gland activity under the control of the sympathetic branch of the autonomic nervous system. The underlying principle is that increased sympathetic activation stimulates the eccrine sweat glands, especially those on the palms and soles, leading to subtle shifts in skin moisture that modulate its electrical properties [113].

It is worth noting that the terminology surrounding EDA can vary considerably across disciplines and historical contexts. It is frequently referred to as skin conductance (SC) or galvanic skin conductance (GSC), while some studies adopt the inverse metric of conductance, thus preferring the use of galvanic skin resistance (GSR). The acronym "GSR" itself is particularly ambiguous, as it has also been widely used to mean "galvanic skin response", a term that may denote either the entire EDA signal or specific components within it [114]. Despite these lexical inconsistencies, all these terms ultimately describe the same core physiological phenomenon: the skin's variable conductivity in response to autonomic activity. For clarity and consistency, we will refer to the complete raw signal acquired for this purpose simply as EDA throughout this work.

### 3.3.1 Sensors

EDA is typically recorded using two electrodes placed on skin regions with a high density of sweat glands, most commonly the fingertips or the palmar side of the hand (see Figure 3.8). The signal itself is generally decomposed into two distinct components: a tonic and a phasic element. The tonic component, also known as the skin conductance level (SCL) or galvanic skin level (GSL), represents slow, baseline fluctuations in conductance over longer periods and is thought to reflect general arousal or sustained mental states. In contrast, the phasic component captures rapid, transient increases in conductance in response to discrete events or stimuli, and is referred to as skin conductance responses (SCR) or, more ambiguously, as galvanic skin response (GSR). While this division is widely accepted, there is currently no standardized approach for isolating these components. Various signal decomposition techniques, ranging from peak detection and filtering to more sophisticated

model-based methods, can yield differing results, complicating cross-study comparisons and interpretation [113].

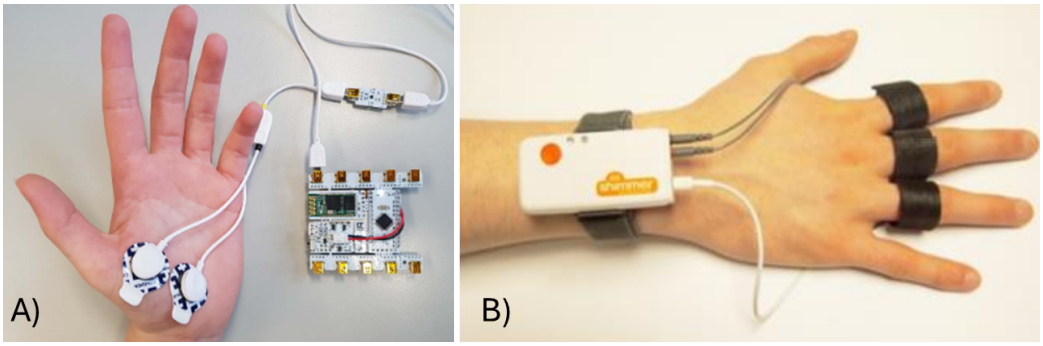


Figure 3.8: Illustrations of different electrodermal activity sensor placements. **A)** Two electrodes positioned on the palm using a BITalino device [115]. **B)** Electrodes attached to the fingers with a Shimmer GSR+ sensor [116].

Yet, all studies presented here used a filter-based technique for signal decomposition, with SCL being considered as the signal under 0.05 Hz and SCR the signal above 0.05 Hz.

### 3.3.2 Features

Typical features extracted from electrodermal activity (EDA) signals are generally organized into two main categories that correspond to the physiological components of the signal: tonic and phasic features [113] (see Figure 3.9).

Tonic features are derived from the SCL and reflect slow-varying changes in baseline conductance over extended periods, typically minutes. These features are used to capture sustained levels of arousal. The most widely used tonic metric is the mean SCL over a given time window, which serves as a marker of general sympathetic activity. Other common descriptors include the standard deviation of SCL, which captures variability in baseline arousal, and the slope or trend over time, which can indicate increasing or decreasing arousal. More advanced statistical descriptors such as skewness, kurtosis, and percentile distributions of SCL can also be employed to characterize long-term fluctuations more comprehensively.

Phasic features, by contrast, focus on SCR peaks, which are typically associated with discrete bursts of sympathetic activity. These peaks can be categorized as either event-related (ER-SCRs), if time-locked to a stimulus,

or non-specific (NS-SCRs), if they occur without identifiable triggers. In stimulus-driven paradigms, ER-SCRs are quantified using features such as peak amplitude (which reflects response magnitude), latency (the delay between stimulus onset and SCR onset), and rise time (time from SCR onset to peak). These metrics offer insight into the intensity and speed of an individual's reaction to stimuli. In non-event-based or resting-state designs, NS-SCRs are commonly analyzed by computing the number of spontaneous SCRs within a predefined time window, which serves as a general index of autonomic arousal.

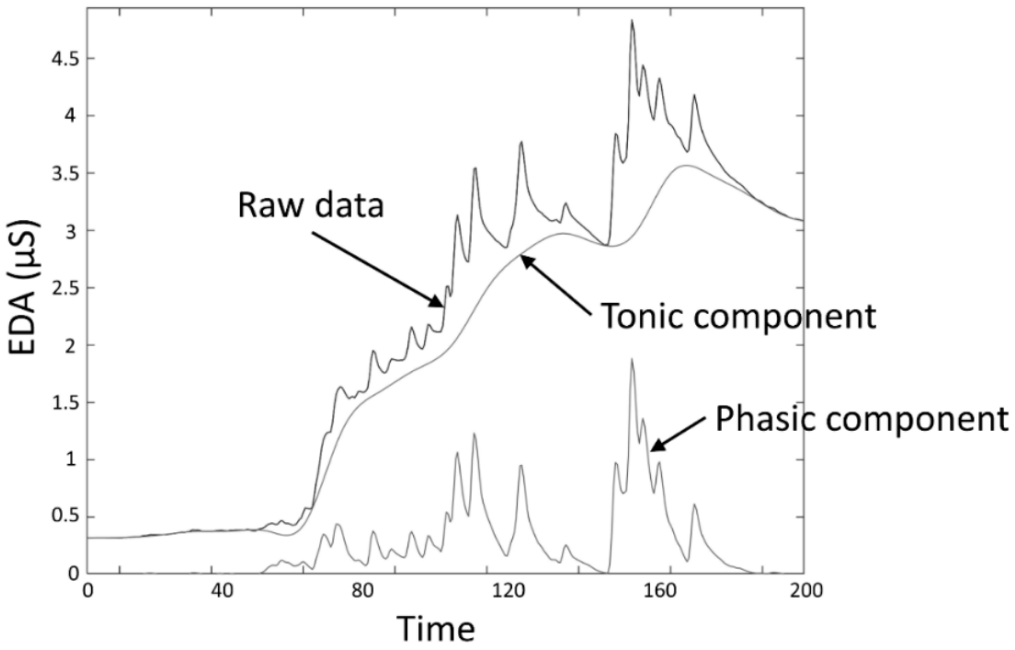


Figure 3.9: Example of a raw EDA signal and its decomposition into tonic and phasic components [113].

In addition to time-domain analyses, spectral methods have been investigated as a technique of EDA analysis [117]. Frequency-domain features aim to quantify the rhythmic components of conductance fluctuations, potentially offering insight into underlying autonomic patterns. While techniques such as power spectral density estimation or wavelet analysis can identify dominant frequencies in EDA signals, their physiological interpretability remains limited and is still an area of ongoing research.

Together, the tonic and phasic features offer complementary perspectives on sympathetic nervous system activity. Their relevance is highly dependent

on the experimental context, particularly the presence or absence of external stimuli, and on the temporal resolution required for the research question at hand.

# Chapter 4

# Mental State Monitoring

Mental State Monitoring (MSM) involves assessing an individual's cognitive and emotional states through the integration of physiological signals, behavioral cues, and subjective inputs. While it originated in clinical contexts, relying primarily on self-reports and observational assessments for diagnosing mental health conditions, MSM has evolved significantly due to advances in neuroscience, psychophysiology, and particularly machine learning. In recent years, the rise of wearable sensors and real-time data processing has extended MSM well beyond clinical settings into operational fields such as human performance monitoring. In this context, MSM enables dynamic and individualized adaptation of systems, such as intelligent user interfaces [118] or neurofeedback training protocols [119]. Such practice can be referred to as MSM, Physiological Computing [49] or passive Brain–Computer Interfaces (pBCIs) [120]. At its core, the principle remains the same: aggregating multiple metrics to derive a composite mental state indicator, either through explicitly defined models or through data-driven approaches (see Figure 4.1). However, the successful deployment of MSM in complex environments, such as Human-Robot Collaboration (HRC), depends heavily on its intended use case. Key design parameters, including the required temporal resolution of mental state estimation, the signal-to-noise ratio of the data, as well as the selection of both the input features and the computational models used for aggregation, all play a critical role in determining system performance and limitations. These challenges and design considerations will be addressed in the following chapter, with a focus on how MSM can be practically implemented to enhance collaboration between humans and robotic systems.

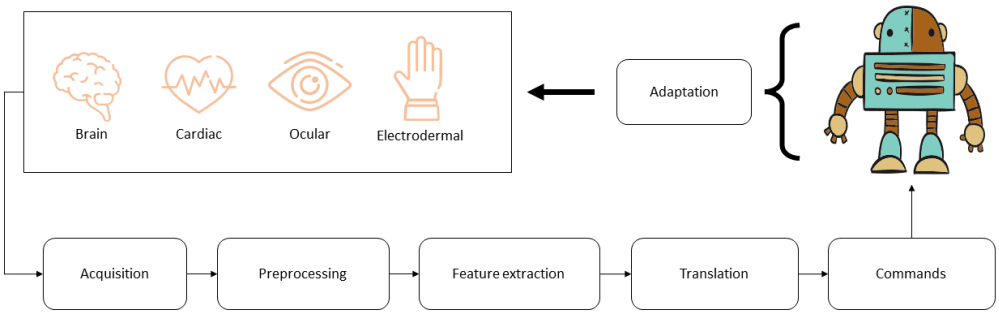


Figure 4.1: Illustration of a physiological loop between human and robot.

## 4.1 Offline and online analyses

The distinction between offline and online analysis, well established in Brain-Computer Interface (BCI) research [121], is equally critical for any MSM system. **Offline analysis** involves processing data after the task was performed, allowing for the use of computationally intensive methods and more degree of freedom when it comes to feature selection. This approach provides flexibility in exploring different algorithms and provides deeper insights into the collected markers of mental states. It helps identify which metrics are most reliable, how they evolve over time, and how they relate to mental states. The outcomes of offline analysis are often used to inform the development of real-time systems.

In contrast, **online analysis** involves processing incoming data in real time to continuously estimate a user's mental state and trigger immediate system responses. This real-time capability allows for dynamic adaptation of interfaces or environments. However, achieving this responsiveness introduces significant technical challenges. Algorithms must be computationally efficient, robust to noise and accurate enough to produce reliable estimates within a sufficiently short time frame. A key requirement of many online systems, especially those based on machine learning, is an initial calibration phase, typically lasting 20 to 30 minutes, during which the system learns the user's baseline physiological responses. This calibration adds to the total duration of the experiment and can reduce system usability, although recent research is exploring methods to minimize or eliminate such requirements [122, 123]. Due to these limitations, and while they allow to characterize the acceptability of the final system with real users, online studies remain significantly less common than offline analyses [121], with real-world implementations being even rarer [124].

A valuable intermediary step between offline and fully online processing is provided by **pseudo-online analysis**, which applies real-time constraints to pre-recorded data streams [125]. Unlike traditional offline approaches that have access to the entire dataset, including future samples, pseudo-online analysis mimics real-time conditions by processing data sequentially, strictly in the order it was originally acquired. This is achieved through fixed-length temporal windows on which are applied denoising, feature extraction, and classification processes, enforcing the same temporal limitations faced in online systems. Although it does not involve live signal acquisition, pseudo-online analyses assess the responsiveness, latency, and robustness of MSM systems under constraints close to those of real-world deployment.

Thus, while online analysis remains an essential final step for creating adaptive MSM systems, the work presented here will mostly focus on offline and pseudo-online analysis. As MSM for HRC is still under development, the current work aims at a deeper characterization of both the mental states of interest and the HRC context.

## 4.2 Signal preprocessing

Signal preprocessing plays a fundamental role in MSM as it aims to enhance the signal-to-noise ratio (SNR), a key metric that reflects the amount of meaningful physiological information relative to background noise. Most physiological signals used in MSM, such as EEG, ECG, or EDA, are inherently low in amplitude and susceptible to various forms of interference [126] (see Figure 4.1). In real-world environments, where users move freely and conditions are less controlled, noise contamination increases significantly [78]. Without adequate preprocessing, these disturbances can obscure the underlying physiological patterns needed for accurate mental state estimation. Moreover, the clearer the signal, the more efficiently mental estimation models can extract relevant features and produce reliable outputs. Thus, robust preprocessing is not just a technical requirement, but a cornerstone of any reliable MSM pipeline.

| Origin | Source | Examples | Affects |
|---|---|---|---|
| Internal or Physiological | Ocular | Eye blinks, flutter, saccades | EEG |
| | Cardiac | Heartbeat, pulse-related activity | EEG |
| | Muscular | Chewing, swallowing, clenching talking, eyebrow movements | EEG, ECG |
| | Respiratory | Electrode shifts from chest expansion or breathing | EEG, ECG |
| | Sweat | Changes in electrode conductivity due to inspiration | EEG, ECG, EMG |
| External or Environmental | Instrumental | Electrode displacement, cable movement, poor grounding | EEG, ECG, EMG |
| | Interference | Electrical noise, optical cross-talk, sound, magnetic fields | EEG, ECG |
| | Motion | Head or body movement, tremor walking, running | EEG, ECG, EMG, PPG |
| | Ambient | Noise, air drafts, temperature shifts, stress-induced changes | ECG |

Table 4.1: Common artifacts affecting physiological signal acquisition, categorized by origin, source, and typical impact reprocuded from the original paper [126].

However, identifying what constitutes "noise" is not always straightforward. Some physiological artifacts, such as eye blinks in EEG data, may traditionally be considered undesirable because they distort the underlying brain signal [127]. Yet they can carry contextually relevant information and be used as valuable features for mental state inference [128], as eye blink rate can reflect fatigue or cognitive effort for example. This highlights a growing understanding that the distinction between artifact and signal is often context-dependent. Adding to the complexity, recent studies in EEG research have shown that sophisticated preprocessing pipelines may have limited impact under certain conditions. Specifically, it was reported little to no difference in classification outcomes between raw and preprocessed EEG data in controlled lab settings with high-grade equipment [129]. While these findings may not generalize well to wearable or real-world settings, they challenge the necessity of current preprocessing techniques and point to a possible need for new solutions.

Despite these nuances, motion artifacts and power line interference constitute common sources of signal degradation across physiological signals. Motion artifacts, caused by voluntary or involuntary body movement, intro-

duce abrupt and high-amplitude deviations that can mimic or obscure physiological patterns. While these can often be minimized in laboratory settings through strict participant instructions, such restrictions are impractical in naturalistic applications. Strategies to address them include techniques such as adaptive filtering [130, 131], wavelet transform [132, 133] and blind source separation (BSS) [134], and can go up to excluding a portion of the data. In contrast, line noise, typically from the power supply and surrounding electronic equipment at 50 or 60 Hz, is usually handled with notch filters.

Each physiological modality also presents its own specific noise challenges. In EEG, two dominant artifacts are ocular movements (blinks and saccades) and electromyographic activity from facial muscles. These often overlap with the same frequency bands used to infer cognitive states. Blind source separation methods [135], spatial filtering [136] and channel regression [137] are standard tools for dealing with these artifacts. In ECG and PPG, baseline wandering from respiration or posture changes can obscure heart rate variability measures. High-pass filters [138] and decomposition methods [139] help mitigate these effects while preserving key features like R-peaks. PPG is relatively unaffected by other factors than motion, except for ambient light interference which is often managed directly at the sensor level. EDA signals suffer from slow-drift artifacts due to skin hydration changes, temperature shifts, and variable skin-electrode impedance. These are often addressed using low-pass filters and signal smoothing techniques [113].

It is also worth noting that preprocessing pipelines are often tailored to the intended use case: offline analysis allows for more extensive corrections while real-time applications require lightweight, causal processing methods that maintain low latency.

In summary, signal preprocessing remains a critical yet evolving aspect of MSM. While traditional methods like filtering and artifact rejection continue to serve as foundational tools, ongoing research is challenging assumptions about their effectiveness, especially under ideal lab conditions. As the field moves toward more ecologically valid settings, the development of more adaptive, context-aware preprocessing pipelines will be essential to bridge the gap between theoretical importance and practical reliability.

## 4.3 Relevant features

Once signal preprocessing has improved the SNR, the next critical step in MSM is the extraction of features that can inform classification [140, 141]. Approaches to feature selection typically fall into two categories. Data-driven strategies rely on feeding a large pool of features into automated methods,

which then identify patterns associated with the mental state of interest [142]. This approach offers flexibility and the potential to uncover novel markers, but it comes at the cost of interpretability and generalizability, particularly when noisy or redundant features are included. In contrast, a priori selection restricts the analysis to features that have already been shown to correlate with specific mental states [143]. While this strategy often yields more robust and interpretable models, it risks missing task-specific information. While both approaches are popular in MSM, the choice between them typically depends on the application goals and constraints.

Although physiological signals lie at the core of MSM research, offering rich information and the highest temporal resolution, subjective and behavioral measures remain indispensable complements. For a long time, and still today, studies of human–robot collaboration have relied heavily on questionnaires and performance metrics to assess user experience and task outcomes. However, in the absence of physiological data, these measures rarely map cleanly onto specific mental states. Instead, they are often interpreted at a broader level, such as overall interaction quality. To avoid this ambiguity, the present review emphasizes metrics that can be explicitly linked to cognitive states. Consequently, it reports only those subjective and behavioral features that have been directly associated with a particular mental state and, in most cases, employed alongside physiological measures to provide converging evidence.

This section surveys existing HRC studies that employ features associated with the three cognitive mental states of particular relevance in the field: cognitive effort, stress, and error detection. Because the usefulness of these features is highly context-dependent, and the temporal resolution required varies across experimental settings, this review considers subjective, behavioral, and physiological measures alike. The interdisciplinary nature of MSM research further complicates the picture, as terminology and theoretical frameworks differ widely between studies. To address this challenge, findings are organized here by experimental design rather than by the mental states explicitly reported. While this choice limits the ability to draw firm conclusions about the underlying constructs, it provides a clearer and more consistent basis for comparing the features leveraged across different contexts.

### 4.3.1 Cognitive effort

This subsection reviews studies in which humans collaborate with robots on tasks designed to vary in difficulty (see Figure 4.2). When considered, a real-time robotic adaptation using this mental state focuses on adjusting the

robot's pace [53] or dynamically reallocating tasks between human and robot [144, 145].

Across studies, the targeted mental state is referred to using various terms, including workload [146, 54, 57], cognitive load [53], stress [58], and fatigue [145, 144]. Despite this terminological variability, the analysis in these studies always focuses on differences between difficulty levels, i.e., task demands. Consequently, these studies seem to reflect what is defined in this work as cognitive effort.

To estimate this mental state, researchers combine subjective, behavioral, and physiological metrics. Subjective data are typically collected using standardized questionnaires, with a vast preference for the NASA-TLX questionnaire [146, 54, 57, 53, 145]. Only one study additionally refers to a 9-point rating scale [53]. Behavioral indicators such as task completion time [57] and number of errors are also widely used [57, 58, 54] to evaluate task performance.

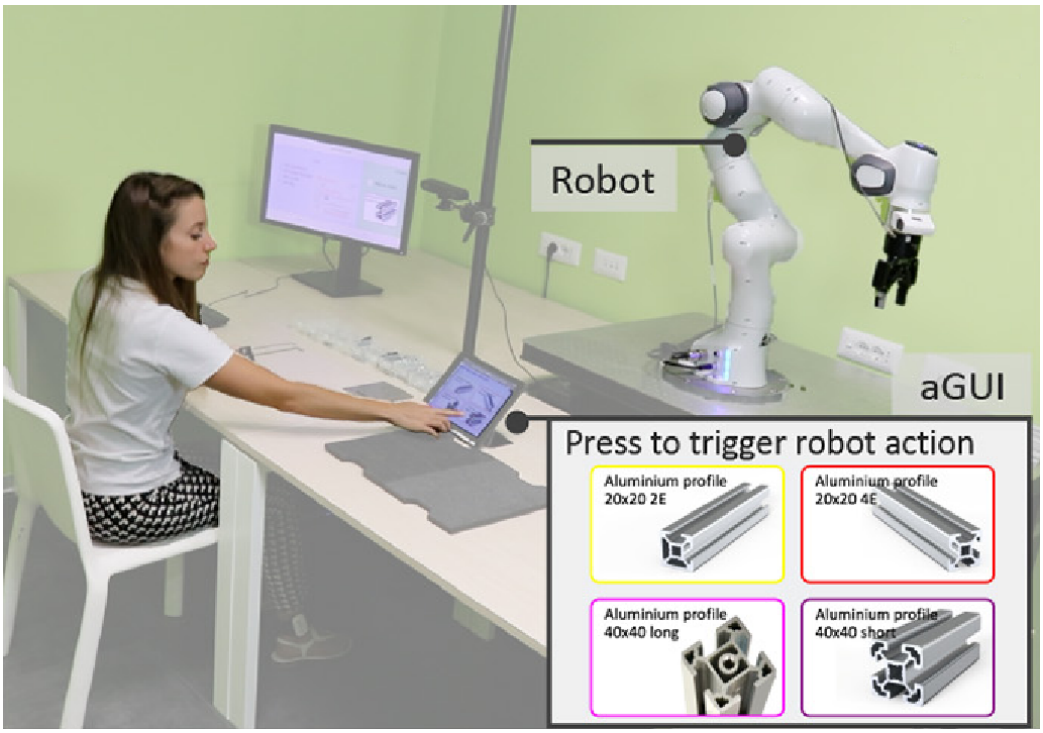


Figure 4.2: Illustration of a collaborative experiment in which a participant's cognitive effort is monitored while performing an assembly task under varying levels of robotic assistance [146].

Among physiological data, EEG has been used with data-driven approaches [54, 53] or frequency-based features, mainly with alpha and low-gamma band power from frontal regions [57]. Cardiac features, including RR intervals from chest straps, wristbands, or smartwatches [145, 144], as well as the LF/HF ratio from chest straps [146], have also been employed. EDA features like the SCR mean peak amplitude have been reported [58].

In conclusion, although cognitive effort has been widely explored using physiology in domains such as human factors and neuroscience, its investigation within HRC remains relatively sparse. While existing studies support the relevance of classical physiological markers, like EEG band power and heart rate variability, the limited number of HRC-specific experiments constrains the robustness and generalizability of these results. The lack of standardized experimental protocols and publicly shared datasets further impedes cross-study comparisons and model reproducibility. An exception, however, lies in the frequent use of the NASA-TLX questionnaire, which provides a consistent offline measurement of perceived workload across studies. Nonetheless, the field still lacks a solid foundation for reliably identifying the most informative features for cognitive effort monitoring in real-world HRC scenarios.

### 4.3.2 Stress

This subsection reviews studies where humans and robots share a physical environment but do not collaborate on an explicit task. These scenarios typically involve the robot moving within a shared space, and the primary objective is to understand how different aspects of the robot's behavior—such as speed, trajectory or proximity—impact the human's stress response. The ultimate goal is often to inform the design of robot behaviors that are perceived as socially acceptable or emotionally comfortable (see Figure 4.3). Among the reviewed works, only one study explores the potential for real-time robot adaptation based on the individual's physiological stress profile [147].

Stress is the most commonly reported mental state in this context [146, 56, 148, 63, 62], although it is sometimes framed more broadly as an affective response [147].

Subjective measures are widely used in these studies to assess participants' internal states. These include validated questionnaires such as the Semantic Differential questionnaire (SD) [148], the Self-Assessment Manikin (SAM) [56], and the Negative Attitude toward Robots Scale (NARS) [56], as well as custom scales replicated from previous studies [56, 146] or simple numeric rating scales [63, 62]. Yet, no questionnaire seems more popular than others. Unlike the cognitive effort studies, behavioral metrics are not used

to complement physiological signals in the works considered. This is likely due to the passive nature of the scenarios, which do not involve goal-directed metrics.




Figure 4.3: Illustration of an experimental scenario in which a participant's performance is monitored while performing a task under two distinct robot interaction styles: an encouraging robot versus a stress-inducing robot [62].

Physiological signals, especially EDA, play a central role in assessing stress. Features such as SCR peak amplitude [146, 56, 148, 147, 63, 62] and mean SCL [146, 62] are frequently used. HR collected via ECG has also been employed in one study [147].

In summary, while stress-related studies are crucial for advancing HRC, they tend to focus more on human-robot co-presence than on active collaboration. This type of interaction typically allows for lower temporal resolution in mental state monitoring. As a result, these studies often rely on slower physiological signals such as EDA and post-interaction questionnaires.

### 4.3.3 Error detection

This subsection focuses on studies where physiological signals are used to detect human responses to unexpected or erroneous actions performed by a robot. In these scenarios, the goal is not to monitor a continuous mental state but to identify discrete moments when the human perceives a robot error. A real-time adaptation would use such information to trigger behavioral corrections in the robot [68, 149] (see Figure 4.4) or to guide learning algorithms, such as reinforcement learning, in refining robot behavior [70].

In all of the reviewed studies, the primary focus is on detecting error-related potentials (ErrPs) from EEG signals. This particular ERP is characterized by two distinctive components: a negative peak around 200 milliseconds after the event, often referred to as the error-related negativity (ERN or N200), and a following positive peak at approximately 300 milliseconds, known as the P300. Together, these components reflect the human brain's automatic recognition of a discrepancy between expected and observed outcomes.

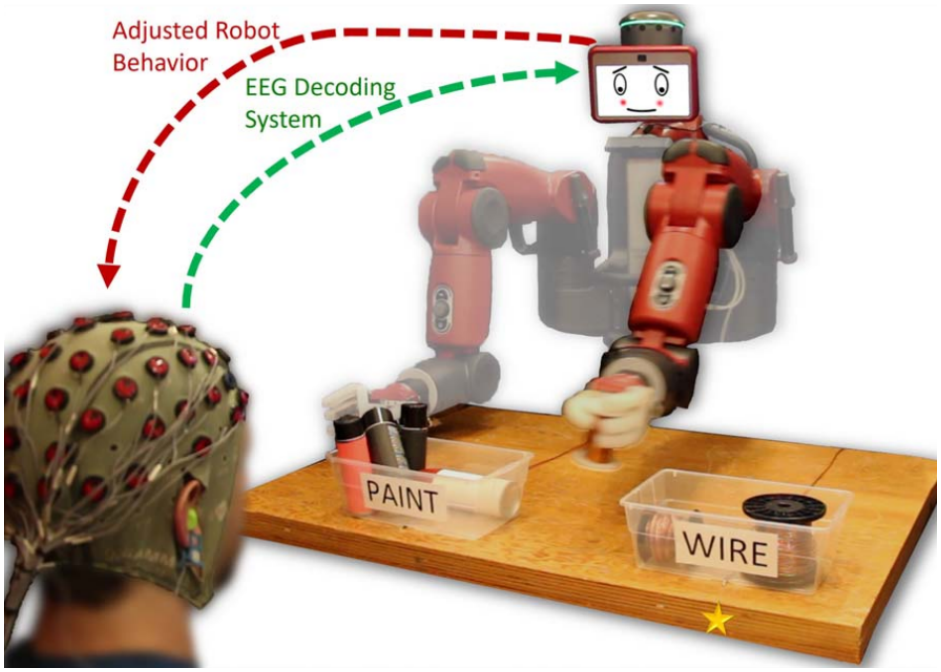


Figure 4.4: Illustration of a binary-choice scenario in which the errors performed by the robot are corrected online by the participant's physiological responses [68].

Unlike the other mental states reviewed in this work, these studies do not

incorporate subjective self-reports, likely because ErrP detection is event-focused and time-locked to stimuli, making introspective ratings less practical or informative. Similarly, only one study report behavioral metrics [67], using reaction time and number of errors.

Physiologically, all reviewed studies rely heavily on EEG data, as accurately detecting ErrPs requires capturing both the temporal dynamics and spatial distribution of the activity. Most studies adopted a data-driven strategy, feeding time-locked EEG signals from all available channels directly into classification models without prior feature extraction [70, 67, 149]. To address this high dimensionality, two studies implemented dimensionality reduction techniques: principal component analysis (PCA) [149] and XDawn spatial filtering [70]. Additionally, one study opted for a more targeted strategy by calculating covariance matrices from a subset of frontal EEG channels [68], leveraging the well-established role of frontal regions in error detection and performance monitoring.

In summary, monitoring error detection in HRC is characterized by its demand for high temporal precision, which limits the choice of input modalities almost exclusively to EEG. Unlike other mental states, neither subjective reports nor behavioral indicators are used to complement physiological signals in these studies. In addition, despite the well-established theoretical shape of ErrPs, most research adopts data-driven approaches rather than template matching, emphasizing the need for models that can accommodate the variability found in real-world conditions. This highlights both the specificity and complexity of error detection in HRC, where robust, temporally precise neural features are essential.

## 4.4 Classification

Once relevant features have been extracted, the next step in MSM involves mapping these features to discrete or continuous estimates of mental states. This process, often referred to as classification or regression depending on the output format, can be addressed using a wide spectrum of methods, ranging from simple rule-based systems to complex data-driven models.

The most basic approach relies on manually defined thresholds derived from prior literature. These rule-based systems are appreciated for their transparency, ease of implementation, and suitability for real-time use. For example, reduced heart rate variability or changes in EEG spectral power have been associated with elevated cognitive effort [140, 84], enabling explicit thresholds to be set. However, such methods assume stable physiological markers across individuals and contexts, an assumption rarely met in

practice. Due to the high inter-individual variability in physiological signals [150], threshold-based models often lack robustness and adaptability, limiting their scalability and explaining their decline in favor of machine learning approaches.

To improve adaptability and capture non-linear relationships between features and mental states, many MSM systems employ supervised learning algorithms such as Support Vector Machines (SVM), Random Forests (RF), k-Nearest Neighbors (k-NN), or Linear Discriminant Analysis (LDA) [121]. These models learn data-driven decision boundaries from labeled training examples, allowing for better personalization across individuals. Their relatively low computational cost and relative interpretability make them suitable for both offline and real-time applications. However, they typically require a training phase in online settings, which can be time-consuming and reduce practical usability [151].

A major obstacle for all these approaches lies in the variability of physiological signals. Even within a single session, signals fluctuate due to fatigue, stress, or environmental changes, undermining short-term stability [152]. Across sessions, natural physiological changes, along with habituation and sensor placement differences, are challenging the reuse of trained models and frequently lead to recalibration [153]. Task variability adds a further layer of complexity: even when designed to elicit the same mental state, task modality, structure, and context can modulate how physiological markers manifest [154]. To address these issues, research has increasingly explored transfer learning, which seeks to design models that generalize across individuals, sessions, and tasks, thereby reducing or eliminating the need for recalibration [155, 156].

In parallel, deep learning has become increasingly prominent in MSM, especially for processing high-dimensional and temporally complex data such as EEG. Models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrids such as CNN-LSTM architectures are capable of learning abstract spatiotemporal representations directly from raw or minimally processed signals [121, 157]. This end-to-end learning pipeline can effectively merge features from multiple physiological modalities [158] and has shown promise for reducing sensitivity to physiological variability [159, 160], making it an appealing candidate for MSM. However, deep models often require large, well-labeled datasets to avoid overfitting and are typically opaque in their decision-making, complicating interpretability. Furthermore, their computational demands make real-time deployment challenging. As such, while deep learning models may yield strong offline performance with the appropriate dataset, their practicality for real-world, real-time MSM remains limited at this stage.

## 4.5 Adaptation

Due to the inherent challenges arising from the interaction between physiological computing and HRC, there are few examples of online robotic adaptation based on mental state monitoring. Nevertheless, given that physiological computing essentially involves modeling human behavior, it holds promise for adaptation. A particularly insightful framework proposes three distinct levels: physical, cognitive, and social levels [9].

At the physical level, robots can modulate the speed and the trajectory during navigation or manipulation tasks. Such adaptations may address safety concerns by modulating features associated with stress and trust. Additionally, they can serve as cognitive effort regulators, especially when the speed of the robot directly influences the pace of the human worker [53].

At the cognitive level, robots can adapt their planning strategies based on the current mental state of humans. For instance, they can dynamically reallocate tasks based on workload or fatigue levels, as seen in mixed-initiative systems [120]. They can also rectify a plan through error detection. In addition, leveraging reinforcement learning techniques enables robots to evolve their decision-making processes over time, attaining a metacognitive level of adaptation. By establishing a feedback loop between humans and robots, physiological computing empowers robots not only to perceive the present mental state of humans but also to gauge the impact of their actions on this mental state, thereby enhancing decision-making capabilities in the long term.

At the social level, robots can adjust their communication strategies based on human mental state. This entails employing various social signals, including non-verbal cues, explicit or redundant information transmission, and even personality adjustments. For example, considering that human preferred interaction styles can vary depending on the task [161], it is plausible to assume that they may also fluctuate based on certain mental states.

## 4.6 Scope of Contribution

This review has highlighted both the promise and the current limitations of using physiological measures to study mental states in human–robot collaboration (HRC). While numerous studies demonstrate the potential of techniques such as EEG, ECG, or EDA to capture aspects of cognitive effort, stress, or error detection, the literature remains sparse, fragmented, and highly heterogeneous. No clear consensus exists regarding which measures should be prioritized, which experimental paradigms are most informative,

or how results can be reliably compared across studies. This lack of common ground has two major consequences: researchers new to the field must often rebuild experimental frameworks from scratch, and even successful experiments produce results that are difficult to generalize or benchmark.

Part of this challenge lies in the diversity of HRC itself. By definition, human–robot collaboration encompasses a broad range of contexts and tasks, each with its own constraints. However, this should not prevent the establishment of baseline paradigms. Psychology and neuroscience have long relied on standardized, fundamental tasks to probe mental processes under controlled conditions, before moving on to more complex and applied settings. A similar effort is necessary in HRC to enable cumulative progress, reproducibility, and meaningful comparison between studies.

In this light, the present PhD aims to contribute to the development of such a foundation by systematically analyzing how physiological signals can be integrated into HRC research and by critically assessing their strengths and limitations. It first investigates electrophysiological markers of cognitive effort with real-world constraints, explicitly accounting for challenges brought by inter-individual variability, wearable sensors, and the integration of a robot in the loop. Building on this analysis, it introduces several experimental paradigm inspired from neuropsychology to study human information processing in HRC from a fundamental perspective, while offering ready-to-use tasks and a publicly available dataset to facilitate replication and comparison across studies. Finally, it complements this focus on cognitive effort with investigations of error detection in HRC, providing a broader view of how different mental states can be monitored through physiological measures.

# Part II

# Contributions

# Chapter 5

# Make it work

## 5.1 Motivation

Mental state monitoring (MSM) for human-robot collaboration (HRC) comes with several constraints, from hardware to software level. Medical-grade sensors, while offering high signal-to-noise ratio (SNR) and superior accuracy, impose movement constraints and require extensive setup, limiting their usability in dynamic environments. Hence, wearable sensors are more suitable for real-world applications due to their portability and ease of use, but typically exhibit a lower SNR, making them more susceptible to environmental noise. This challenge is especially pronounced in EEG-based mental state monitoring, where the differences between wet and dry electrodes significantly impact both signal quality and classification performance (see Section 3.1.1).

Beyond sensor-related challenges, real-world constraints also impose limitations on data processing and classification strategies (see Section 4.2). Online monitoring systems must prioritize epoched signal analysis to ensure timely feedback while also considering participant comfort by minimizing the number of trials used for training. These constraints inherently reduce the amount of available data, limiting the performance of machine learning models.

Finally, conducting multiple sessions is of great importance in HRC, as brief interactions can potentially bias participant behavior and their perception of the robot, limiting the ecological validity of the findings. This concern extends to Brain-Computer Interfaces (BCI), for which inter-session variability of neurophysiological signals poses a significant obstacle to system robustness and long-term usability (see Section 4.4).

Understanding the performance gap between ideal laboratory conditions

and the complexities of real-world HRC, such as user movement, environmental noise, real-time feedback requirements, and temporal variability, is essential for identifying the key challenges that still hinder progress in the field. This would be a critical step toward the development of physiological monitoring systems that are not only robust, accurate, and responsive, but also truly applicable and effective in dynamic, real-world collaborative scenarios.

Therefore, the objective of this first experiment was to evaluate the real-world usability of electrophysiological features for detecting changes in cognitive effort during an HRC task. Specifically, it sought to answer the following research questions:

- Do highly wearable sensors provide signal of sufficient quality for mental state monitoring during HRC?
- To what extent does inter-session variability influence these physiological signals and the reliability of mental state estimation?

The following hypotheses were made regarding the metrics:

H1 – Cognitive effort will impact subjective, behavioral and physiological metrics, i.e.:

  (a) Self reported effort, response time, HR, EEG engagement ratio and mean phasic EDA will increase along with cognitive effort;

  (b) Answer accuracy and HRV will decrease when cognitive effort increases;

H2 – These effects of cognitive effort will vary between session.

H3 – Machine learning-based classification accuracy will exceed the adjusted chance level for the within session case, but drop below for the between session case (because of H2).

## 5.2 Overview

To fulfill the objectives of this study, a two-session experimental protocol was implemented involving the PR2 humanoid robot and an object manipulation task coupled with a digit span paradigm. This memory task enabled the controlled induction of varying levels of cognitive effort while participants interacted with the robot.

During both sessions, participants' physiological signals were recorded using highly wearable sensors without movement restriction. These recordings were complemented by behavioral performance metrics and subjective feedback to provide a deeper understanding of the participants' cognitive states.

The setup was specifically designed to examine how real-worl constraints-such as the physical presence of a humanoid robot, the use of mobile acquisition devices, and variability across repeated sessions—affect the reliability and accuracy of physiological monitoring in HRC.

The chapter begins with a detailed description of the experimental design followed by information about data collection and extraction during the user study. Classical statistical analyses are then presented to evaluate effects of task difficulty and session on the metrics. Then, a focus on robot-related artifacts highlights the need for careful preprocessing. Finally, MSM pipelines are proposed for each signal and implications of these first results for real-world MSM in HRC are discussed.

## 5.3 Methods

### 5.3.1 Robot

The Personal Robot 2 (PR2, see Figure 5.1) is a standard robotic plateform used for HRC research. Because this experimental campaign focuses on the PR2's ability to grasp and manipulate objects in human environments, the robot was only used in a static position. In addition, its speakers were used to explicitly sequence the interaction with audio cues using text-to-speech Google API.

### 5.3.2 Physiological Sensors

Electroencephalography (EEG) measures brain electrical activity with a very high temporal resolution (i.e. ms). EEG data were collected using a 20-channel Enobio (Neuroelectrics) with dry electrodes placed on a cap according to the 10/20 international system and sampled at 500 Hz. Electrocardiography (ECG), photoplethysmography (PPG), EDA sensor and eye-tracking, respectively measure cardiac activity, blood volume changes, electrodermal activity and ocular behavior. ECG data sampled at 125 Hz were collected using a Faros 360°. EDA and PPG data were both collected using a Shimmer3 GSR+ Unit and sampled at 50 Hz.

### 5.3.3 Software

Recording time series through multiple devices raises many issues, and particularly synchronization issues. The open-source ecosystem lab streaming layer (LSL) [162] was used to unify the recordings, both signals and markers. Although this ecosystem allows for real-time signal processing, only offline processing was performed in this study. Hence, LabRecorder allows to save the synchronized time series in a file following the open source xdf standards [163].

### 5.3.4 Tasks

In order to induce various cognitive effort levels, a digit span task was used. This typical memory-load task has already been applied with success to robotic tasks monitored via EEG [53]. Here, the digits were displayed sequentially on a screen located on the wall near the robot. Sequences of 1, 3 and 7 digits were expected to result respectively in a low, medium and high cognitive effort.

The joint manipulation task itself consisted in piling cubes with a robot and began with the participants seated at a table and the robot (PR2) standing on the other side, facing them in its rest position (see Figure 5.1). An audio cue indicated that PR2's turn was beginning ("My Turn"). During this turn, PR2 picked the cube at its right with its right hand and placed it at the center of the table. This sequence of actions was entirely scripted in order to reproduce exactly the same behaviour and the same duration (25 seconds) for each trial. Then, another audio cue indicated that participant's turn was beginning ("Your Turn"). During this turn, participants were instructed to place, with their right hand, the cube at their right on the top of the cube placed by the robot.

In addition, some behaviors (i.e. congruent or incongruent head movements) were added to elicit surprise but did not succeed to induce it. These incongruences were scripted and occurred a fixed amount of times.

Lastly, the participants underwent a resting state period of 5 minutes before the main task.

Regarding the scripts and anticipating the needs for subsequent data analysis, several markers were generated during these tasks, allowing to process physiological signals over various time-windows.

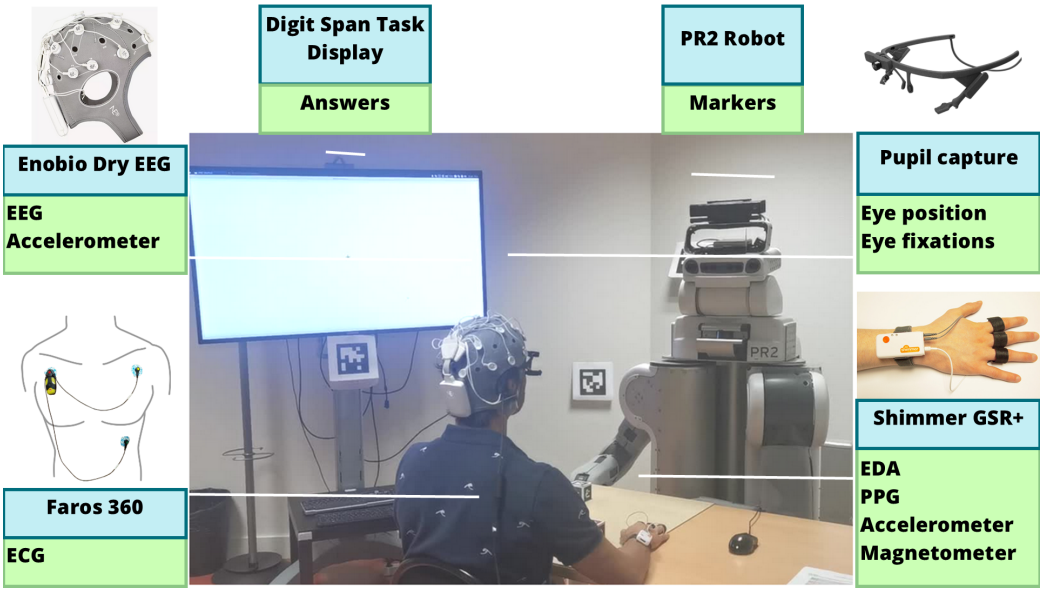


Figure 5.1: PR2 moving the cube in front of the participant during the cube piling task. All gathered signals (green rectangles) are listed below their source (blue rectangles).

## 5.4 User Study

### 5.4.1 Participants

Thirty volunteers took part in the experiment. Four participants' data were excluded due to technical problems. Five others where excluded due to poor signal quality. The remaining 21 participants (11 females, 10 males) had an average age of 28 $\pm$ 5.9 years. Prior knowledge with robots scored 2.1 $\pm$ 1.1 on a scale of 1 "not at all" to 5 "very much". On the same scale, prior knowledge with physiological sensors scored 2.1 $\pm$ 1.5. The participants were instructed about the experimental protocol and provided an informed consent. The study was approved by the institutional ethics review board of the University of Toulouse (project n°2022_525).

### 5.4.2 Protocol

The experiment was scheduled across two sessions which occurred one week apart, at the same time of day. As detailed in Figure 5.2, upon arrival at the laboratory for their first session, participants were informed about the purpose and the procedure of the study. The researchers asked if there were

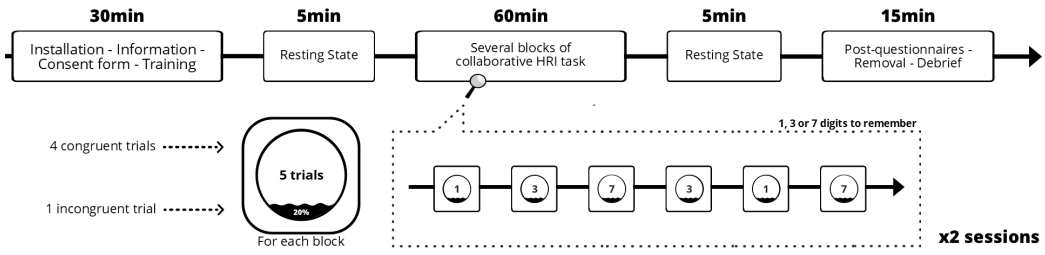


Figure 5.2: Overview of the experimental protocol, including the global timeline (top), the structure of a block (bottom-left corner) and an example of the order that blocks can follow (bottom-right corner). Congruency of trials was linked to scheduled and scripted unexpected robot behavior, which is not analyzed here.

any unclear questions or if the participants wanted to withdraw at this point. Then, they were asked to sign the informed consent form, the form for the inclusion and exclusion criteria, the demographic's questionnaire and the Edinburgh handedness inventory [164].

Next, participants were introduced to the task and the robot (see Figure 5.1). A short training sequence for the task composed of six trials (3 in the low cognitive effort condition and 3 in the high cognitive effort condition) and which lasted approximately 5 minutes was provided. Among these trials, only one (in the low cognitive effort condition) belonged to the incongruent condition in order to check participant's reaction while maintaining this condition infrequent. Participants were free to ask questions at any time and the researcher ensured that the task was understood by the participants.

After completion of the training phase, the physiological sensors (ECG, EEG, EDA, PPG) were set up. Then, participants completed a resting state period during which they were instructed to do nothing. Once the resting state and the Amsterdam Resting State Questionnaire (or ARSQ, that allows to study the link between ongoing brain activity and cognition [165]) were completed, the participants began the experimental task. This task lasted approximately one hour and was composed of 60 trials gathered in 12 blocks of 5 (see Figure 5.2). Each trial began with an audio cue ("New trial") followed by a pseudo random sequence of digits displayed on a screen. Next, they had to remember this sequence while performing the joint manipulation task. Finally, they had to recall the digits in the correct order by typing them on a virtual keyboard.

All trials of a given block belonged to the same cognitive effort condition, thus the cognitive effort condition only varied between blocks. In addition,

one out of the five trials of each block was incongruent (robot head moved incongruently when it moved the cube). These blocks were completed by the participants in a pseudo-randomized order.

Between each block the participants completed the Instantaneous Self Assessment questionnaire (ISA) [166, 167] and gave their subjective feedback on how natural the behaviour of the robot appeared to them (on a 5-point likert scale), they were also able to take a short break if needed. The Human–Robot Interaction Evaluation Scale (or HRIES, an approach of anthropomorphism in HRC through four components: Sociability, Disturbance, Agency and Animacy [29]) was only completed once, at the end of the 12 blocks, to capture their subjective perception of the task and the robot.

At least two researchers remained present at all times, outside of the participant's field of view, to monitor the task and ensure that each trial was completed accurately. One was in charge of the well-being of the participants, paying particular attention to any discomfort that could occur because of the sensors. The other was carefully monitoring the robot, both on the computer and in the room, ready to stop it in case of an unplanned behaviour or if requested by the participant.

Next, the participants completed another five-minute resting state period. Finally, the researchers removed all the sensors and asked participants if they had any questions regarding the research, informing them that, should any questions arise, they could contact the researchers via the contact information given on the information sheet. Participants were also financially compensated for their participation in the study.

The second session followed the same procedure but omitted the consent form as well as the demographics questionnaires and Edinburgh Handedness Inventory. Including arrival, questionnaires, installation/removal of the sensors and task completion, a usual session lasted two hours.

### 5.4.3 Data preprocessing and feature extraction

Due to technical issues, this study focused on the highly wearable electrophysiological sensors : only dry EEG, PPG and EDA physiological signals were processed during analysis and mental state estimation. The same preprocessing procedure was applied for each trial of each participant (see Figure 5.3). In addition, physiological signals recorded during the task were processed at the epoch level to meet real-time constraints, over the 24 seconds of participant's turn. Only resting-state signal was processed without this constraint.

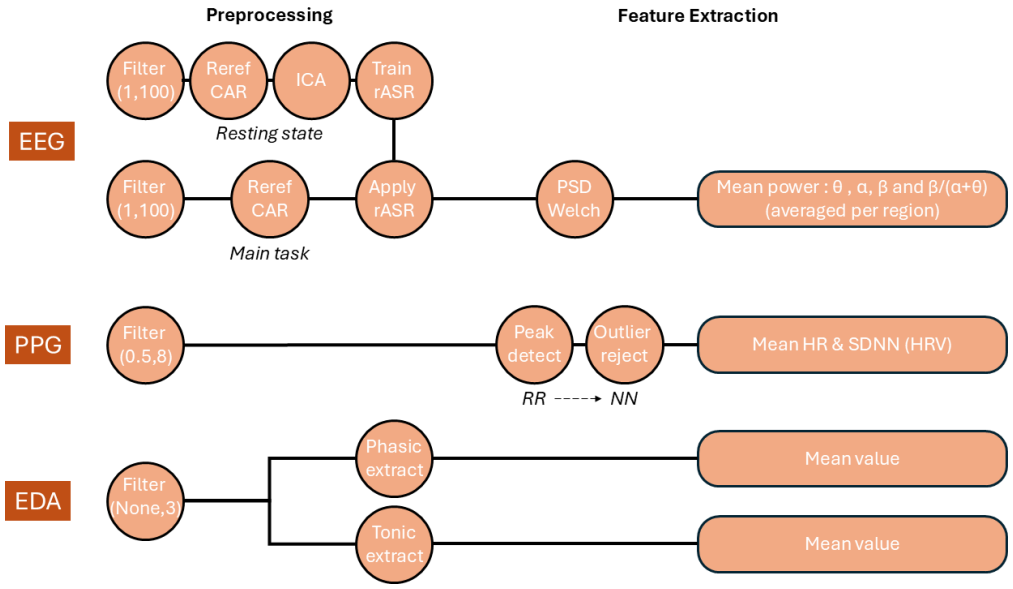


Figure 5.3: Preprocessing pipeline and feature extraction for each signal source. All steps have been conducted using python packages, MNE-python [168] for EEG and PPG, neurokit2 [169] for EDA. Meegkit [170] and pyriemann [171] have also contributed to EEG preprocessing.

Both resting-state and task-related EEG signals underwent a two-step filtering process. First, a finite impulse response (FIR) bandpass filter was applied to isolate frequencies between 1 and 100 Hz. Then, a common average reference (CAR) was used to enhance signal quality. For resting-state only, an additional cleaning step was implemented using Independent Component Analysis (ICA) to remove noise sources. The cleaned resting-state signal was then used to train an artifact rejection algorithm based on the Riemannian variant of the Artifact Subspace Reconstruction (ASR) method [172], which was subsequently applied to task-related EEG signal to mitigate non-neural artifacts in real time. The EEG power spectral density was obtained using the Welch method with non-overlapping windows consisting of 500 data points. The EEG engagement ratio, as defined by Pope et al. [84], was derived from Beta, Alpha, and Theta frequency band power. To improve robustness and interpretability, spectral values were averaged within three anatomically defined regions of interest : frontal, central, and parietal (see Figure 5.4).

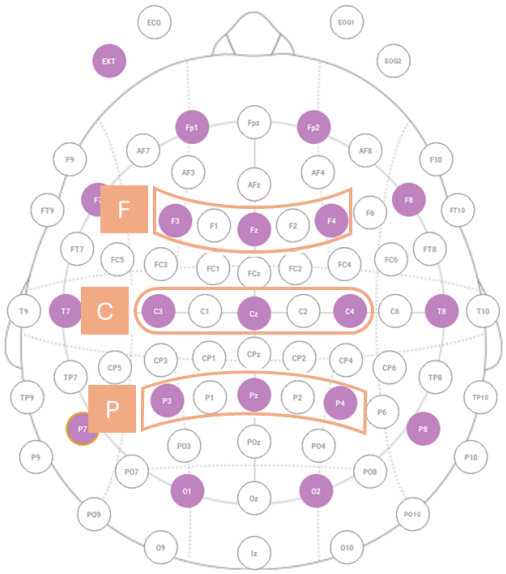


Figure 5.4: Illustration of the selected regions of interest for EEG analysis. The frontal region (F) comprises electrodes F3, Fz, and F4; the central region (C) comprises C3, Cz, and C4; and the parietal region (P) comprises P3, Pz, and P4.

The PPG signal was filtered using neurokit package [169] with a frequency range of 0.5 to 8 Hz. A peak detection algorithm was then employed to identify RR intervals, and those deviating by more than 3 standard deviations from the mean were treated as outliers and excluded. The remaining NN intervals enabled the extraction of heart rate (HR) and SDNN, a measurement of heart rate variability (HRV).

The EDA signal was filtered using neurokit package [169]. Thus, a fourth-order Butterworth low-pass filter at 3 Hz. The processed signal was then decomposed into two components: the tonic component (galvanic skin level, GSL) capturing slow oscillations below 0.05 Hz which reflect baseline electrodermal activity, and the phasic component (galvanic skin response, GSR) representing faster fluctuations between 0.05 and 3 Hz. Statistical analysis was carried out on the mean value of GSL and the mean value of GSR peaks.

## 5.5 Statistical Results

Statistical analysis was performed in an intra-subject design using a two-way repeated measures ANOVA on each metric, with p-values corrected for mul-

tiple comparison using the Greenhouse-Geisser method. While only physiological results are displayed in Figure 5.5, they are reported along with subjective and behavioral results in the following section.

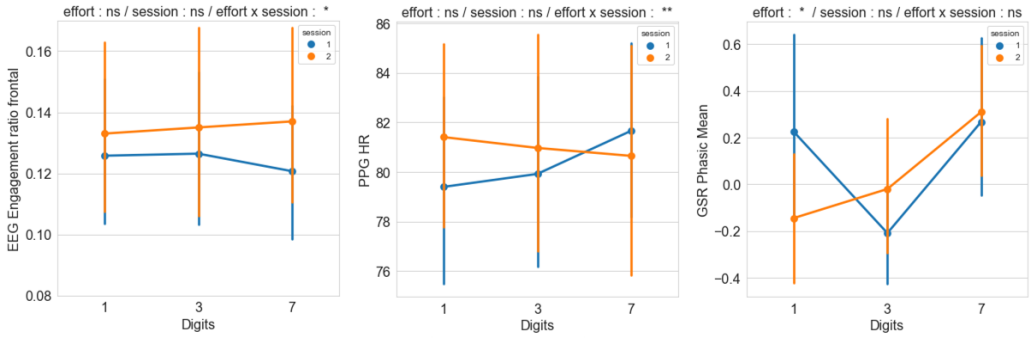


Figure 5.5: Mean value and standard deviation for each metric through conditions and sessions. Effect significativity is reported at the top of each figure with the following rule : ns ($p > 0.05$), * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

### 5.5.1 Sequence length effect

Regardless of the session considered, a significant effect of digit sequence's length was observed for ISA scores ($F(2, 42) = 115.01$, $p < .001$, $\eta_p^2 = .85$) with a mean score of 3.6 for 7-digits sequences compared to 2.2 for 1-digit and 3-digits sequences. The same effect was present for answer accuracy scores ($F(2, 42) = 20.01$, $p < .001$, $\eta_p^2 = .49$) with a mean score of .82 for 7-digits sequences compared to .93 for 3-digits sequences and .95 for 1-digit sequences, as for mean phasic EDA ($F(2, 42) = 3.98$, $p < .05$, $\eta_p^2 = .16$) with a mean value of .29 for 7-digits sequences compared to -.11 for 3-digits sequences and .04 for 1-digit sequences.

### 5.5.2 Session effect

Independently of the digit condition, a significant session effect was observed for Answer accuracy ($F(2, 42) = 7.92$, $p < .05$, $\eta_p^2 = .27$) with a mean score of .89 for session 1 compared to .92 for session 2, and Response time ($F(2, 42) = 19.90$, $p < .001$, $\eta_p^2 = .49$) with a mean value of 1.66 s for session 1 compared to 1.39 s for session 2.

### 5.5.3 Interaction effect

Between sessions, a significant variation of digit sequence's length effect was observed for Answer accuracy ($F(2, 42) = 4.03$, $p < .05$, $\eta_p^2 = .16$), frontal EEG engagement index ($F(2, 42) = 4.20$, $p < .05$, $\eta_p^2 = .17$), Heart Rate ($F(2, 42) = 9.14$, $p < .01$, $\eta_p^2 = .30$), and Heart Rate Variability ($F(2, 42) = 4.65$, $p < .05$, $\eta_p^2 = .18$). Yet, post-hoc tests only revealed a significant effect for two of them. Answer accuracy in the 7-digit condition increased between sessions ($p < .05$), with a mean score of .79 for session 1 compared to .85 for session 2. Heart Rate during session 1 significantly varied between digit sequence's length, with a higher mean value of 81.66 for the 7-digit condition compared to 79.40 for the 1-digit condition ($p < .05$) and 79.92 for the 3-digit condition ($p < .05$), but not in session 2.

## 5.6 Classifying Cognitive Effort

### 5.6.1 Pipelines

Monitoring mental states requires establishing a robust pipeline, which in turn involves making numerous parameter choices. This section introduces the approach adopted not only in the current study but also in the subsequent ones, ensuring comparability across datasets and providing a broader perspective on how these choices impact classification performance. The key parameters considered for the pipeline include preprocessing, feature domain, data reduction, and classifier type. Yet their relevance varies across signals, for instance EEG preprocessing required carefull design due to its complexity, whereas preprocessing of other signals did not, given its relative simplicity. The following paragraphs detail these parameters for each signal and Table 5.1 provide a summary.

| Signal | Preprocessing | Feature domain | Data reduction | Classifier |
|---|---|---|---|---|
| EEG | Minimal, ASR, ICA | theta, alpha, beta, all, full | None, CSP | SVM, RF, LDA |
| ECG/PPG | | HR, HRV, all | None, PCA | SVM, RF, LDA |
| EDA | | phasic, tonic, raw, all | None, PCA | SVM, RF, LDA |

Table 5.1: Summary of classification pipeline parameters considered for EEG, ECG/PPG, and EDA signals.

### EEG

Preprocessing was especially critical for EEG signals, for which inappropriate parameter choices risk discarding meaningful neural activity alongside noise. To systematically address this challenge, three alternative pipelines were designed and compared. The first, used for statistical analyses and described in 5.4.3, combined band-pass filtering, CAR application and ASR rejection to provide high-quality data while remaining compatible with real-time applications. To assess the impact of artifact correction, a minimal version consisting only of filtering and CAR was implemented, representing the most lightweight option. Finally, a third pipeline incorporated Independent Component Analysis (ICA) on continuous data, followed by automated component labeling and rejection [135]. Although unsuitable for real-time use, this approach was expected to yield the cleanest signals and thus could serve as a quality benchmark.

From the preprocessed EEG signals, covariance matrices were computed and subsequently projected from the Riemannian space into the Euclidean space, where they could be directly used as features for classification. These features were extracted across multiple spectral domains, including the full broadband signal (1–100 Hz), a narrower range encompassing all the canonical bands considered (4–30 Hz), as well as the isolated canonical bands (theta, alpha, beta) to capture frequency-specific dynamics. To further investigate the role of spatial filtering, an additional dimensionality reduction step was tested by applying Common Spatial Patterns (CSP) [91] on covariance matrices, allowing a comparison of classification performance with and without this technique.

### ECG & PPG

For cardiac signals, preprocessing was relatively straightforward compared to EEG. Indeed, artifacts were less pervasive and well-documented pipelines already exist. After filtering and heartbeat detection, interbeat intervals were extracted to compute both heart rate (HR) and heart rate variability (HRV) (see Section 5.4.3 for more details). Multiple HRV metrics were derived from the NN intervals, such as the standard deviation of NN intervals (SDNN) and the root mean square of successive differences (RMSSD), capturing complementary aspects of autonomic regulation.

Feature extraction thus included both HR- and HRV-based descriptors, complemented by descriptive statistics (mean, standard deviation, minimum, maximum, and quartiles) computed directly from the HR series. While the relatively short epoch durations did not allow for robust frequency-domain

analysis, time-domain metrics may provide informative markers of cognitive load. A PCA-based dimensionality reduction step was also tested to compress redundant information across features.

#### EDA

Electrodermal activity was decomposed into tonic, and phasic components to reflect both slow and fast changes in skin conductance. Preprocessing followed established methods, using filter-based approaches to isolate the components (see Section 5.4.3 for more details). As no features more complex than mean value are generally derived from this signal, descriptive statistics (mean, standard deviation, minimum, maximum, and quartiles) were used to characterize its dynamics over each epoch.

Similar to EEG, feature extraction from multiple signal domains was tested, allowing a comparison between raw, tonic, and phasic signal. A PCA step was again used to explore dimensionality reduction, aiming to retain the most informative aspects of the EDA signal while reducing redundancy.

#### Classifiers

The final hyperparameter considered was the choice of classifier. To ensure interpretability and comparability, a consistent set of conventional machine learning algorithms was employed across studies. These included Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), and Random Forests (RF). Deep learning methods and stacking approaches were tested but did not lead to sufficient performances to be considered further.

### 5.6.2 Classification Results

To improve clarity and conciseness, the classification accuracies reported below are limited to the best-performing classifier for each signal under a 5-fold cross-validation, other pipelines will be discussed at the end of the manuscript. Thus, only SVM classifier performances are reported for EEG and Random Forest classifier performances for ECG, PPG and EDA. The figure include the adjusted chance level [173], providing a more comprehensive interpretation of the reported performances.

For EEG features, the highest mean accuracy was achieved using beta band features, reaching 67.77% overall (67.21% in session 1 and 68.33% in session 2). Alpha features followed with 63.22% (62.64% s1, 63.79% s2), while theta features yielded 60.53% (57.35% s1, 63.71% s2). When all frequency bands were included, accuracy reached 62.76% (60.37% s1, 65.15%

s2). In contrast, the "full" band (1–100 Hz) resulted in a substantially lower performance of 55.44% (52.62% s1, 58.26% s2).

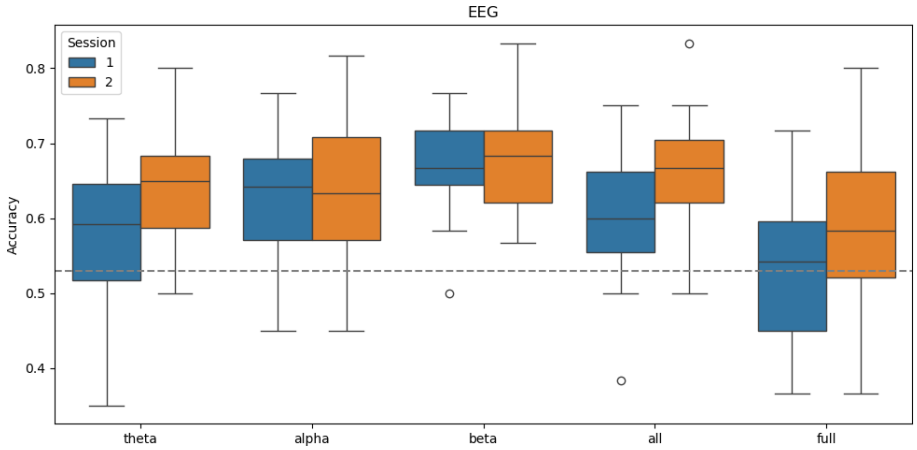


Figure 5.6: 3-class classification accuracy results for each EEG feature. Mean results per session are reported in section 5.6. The dot-line represent the adjusted chance level for alpha = 0.05.

For PPG features, classification accuracy remained close to chance level across all feature sets. The best configuration was obtained using the complete feature set, with 40.08% on average (42.05% s1, 38.11% s2). HR features performed similarly, with 39.55% (41.59% s1, 37.50% s2). HRV features alone gave 39.73% (41.06% s1, 38.41% s2)

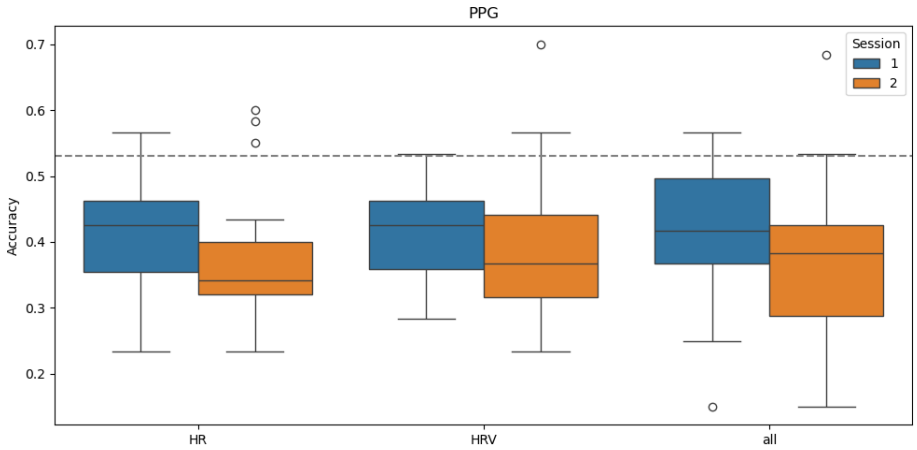


Figure 5.7: 3-class classification accuracy results for each PPG feature. Mean results per session are reported in section 5.6. The dot-line represent the adjusted chance level for alpha = 0.05.

For EDA features, overall performance was slightly above chance but clearly below EEG. The best result was obtained with the full feature set, yielding 55.54% on average (53.55% s1, 57.53% s2). The raw signal feature set provided 53.18% (51.50% s1, 54.86% s2), while tonic features were comparable at 53.04% (51.30% s1, 54.77% s2). In contrast, phasic features performed poorly, with only 38.80% accuracy (38.52% s1, 39.09% s2).

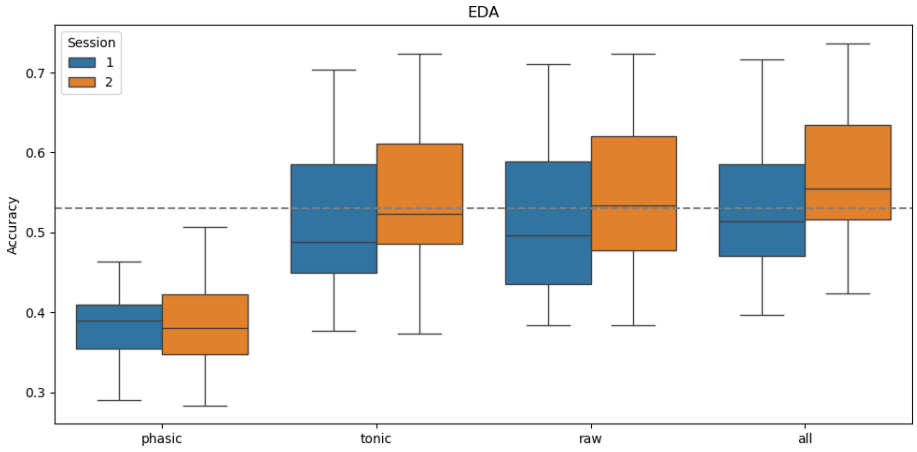


Figure 5.8: 3-class classification accuracy results for each EDA feature. Mean results per session are reported in section 5.6. The dot-line represent the adjusted chance level for alpha = 0.05.

## 5.7 Robot noise

### 5.7.1 Identification

Initial data inspection revealed the presence of strong artifacts that contaminated the raw EEG recordings and could hinder classification performances. To better understand and characterize these disturbances, a dedicated analysis is presented in this section.

As a first step, the experimental protocol was executed without any participant, allowing to record a baseline signal that contained only the artifacts generated by the robot's pick-and-place actions, but no neural activity (Fig. 5.9). This procedure provided a clear reference for isolating and identifying robot-induced noise.

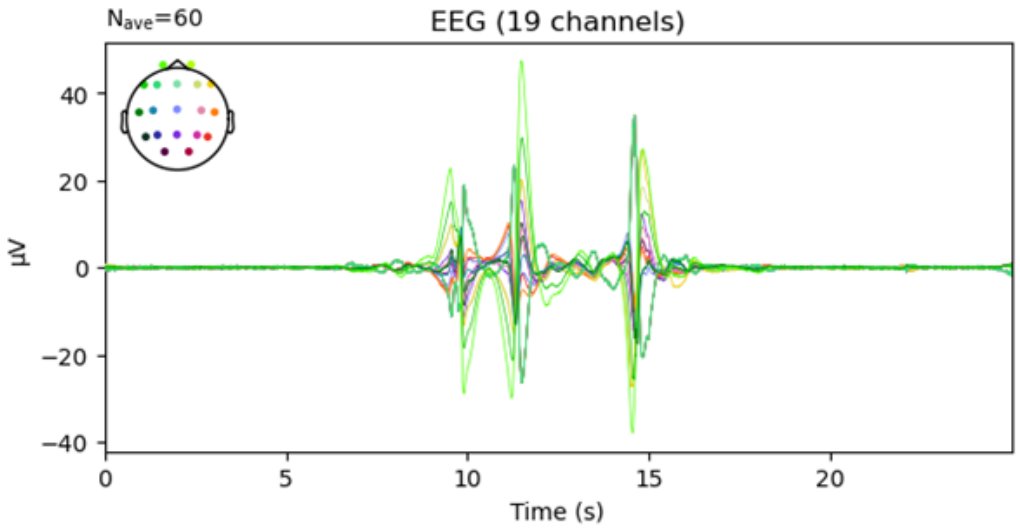


Figure 5.9: Mean EEG signal collected by an empty cap during an experimental session that includes several robot actions (average across trials).

The occurrence of these artifacts strongly aligns with the robot's actions, specifically the opening and closing of its gripper, indicating a direct mechanical origin of the observed noise.

Based on this recording, the 24 seconds used during analysis were split into four windows of 6 seconds each. The first and last windows contained no robot noise while the second window contained two artifacts and the third window contained one artifact.

### 5.7.2 Preprocessing

The EEG data were processed using the lightweight pipeline described in Section 5.6.1, deliberately kept minimal to observe the potential impact of robot-induced artifacts. This preprocessing involved only a band-pass filter and the application of CAR. Following this, covariance matrices were computed and mapped from the Riemannian to the Euclidean space to serve as features. Following the same procedure as in Section 5.6.1, a CSP decomposition was applied, enabling a clearer assessment of how dimensionality reduction could help mitigate such issue.

### 5.7.3 Classification

A classification analysis was performed using Random Forest (RF), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and Minimum Distance to Mean (MDM).

### 5.7.4 Preprocessing Effect

Preprocessing plays a decisive role in EEG-based classification, as even small parameter adjustments can substantially influence performance. In the present analysis, robot-induced noise was addressed by tailoring the frequency bands used for feature extraction. Specifically, EEG features were computed over two ranges: a broad band (1–40 Hz) and a narrower band (4–20 Hz). The restricted band effectively reduced the influence of high-frequency artifacts resulting in more consistent and stable classification performance across analysis windows compared to the broader band.

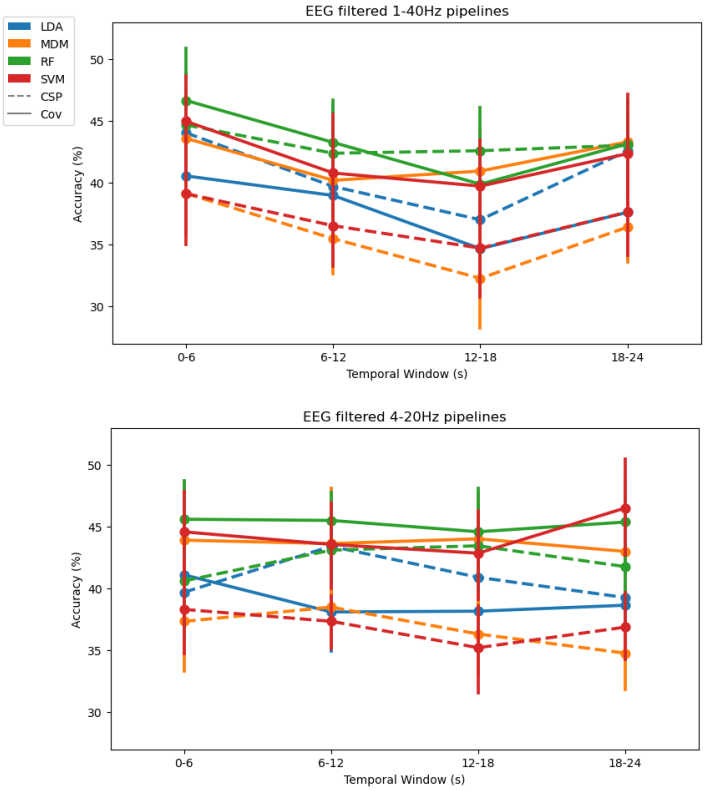


Figure 5.10: Classification accuracies depending on EEG pipelines with different preprocessing steps.

While the classification results of Section 8.4 demonstrate that more advanced preprocessing techniques such as ASR or ICA can address this issue even more effectively, this example tries to illustrate the profound influence preprocessing choices can have on classification performance.

## 5.8 Discussion

The primary objective of this experiment was to evaluate the feasibility of using highly wearable sensors to detect variations in cognitive effort during a human-robot collaboration task. A prerequisite for this goal was to successfully manipulate the participants' cognitive state. Both behavioral and subjective measures confirmed that the protocol effectively induced the targeted mental state: participants reported higher perceived effort with increasing sequence length and performed significantly worse when recalling longer digit sequences, as hypothesized (H1). These findings are in line with

previous research on cognitive effort modulation [53], thereby validating the induction method and supporting the use of the digit span task in HRC contexts.

Physiological data provided additional insights into cognitive effort detection under real-world constraints. Among the signals analyzed, electrodermal activity (EDA)—and particularly its phasic component—proved to be the most reliable, showing consistent modulation with increasing cognitive effort across sessions (H1). In contrast, cardiac activity and EEG responses showed more variability across sessions, highlighting both the potential and the limitations of these modalities. Specifically, while heart rate increased with effort in the first session, this effect vanished in the second, suggesting inter-session variability well-known with physiological signals [153].

Classification results further illustrated the promise and challenges of wearable physiological monitoring in HRC. Intra-session models performed relatively well, with EEG-based classifiers exceeding 60% accuracy and EDA-based models reaching up to 55% accuracy in a three-class setup. These performances, although modest compared to controlled laboratory studies reported in reviews [174], provide a realistic benchmark for what can be achieved in ecological HRC conditions. Importantly, they demonstrate that EEG and EDA carry discriminative information about cognitive effort, partially validating our hypothesis (H3). However, inter-session generalization remained problematic: classifiers trained on one session often failed when applied to another. This degradation highlights the instability of physiological features across time and supports H2 regarding the influence of session-related variability. It also emphasizes the need for adaptive approaches such as transfer learning, domain adaptation, or per-session calibration [155, 156, 159] to make physiological state monitoring viable in long-term, real-world deployments.

Along with physiological responses, both response time and answer accuracy exhibited significant variations across sessions, indicating that behavioral measures can also vary across time, as previously suggested [175, 176]. This phenomenon could be explained by various factors including a habituation effect and a variation in engagement between sessions.

It's noteworthy that answer accuracy allows both to discriminate cognitive effort across sessions and is impacted by the session effect. This suggests that, while answer accuracy may be smoothed across sessions, the variation across difficulty conditions remains substantial enough to remain significant. However, when it comes to physiological measurement, the effect size does not seem to be sufficiently high to withstand the effect of session, as is generally observed in BCI [153].

These findings highlight the importance of carefully balancing signal ro-

bustness and informativeness when selecting modalities for cognitive state monitoring. While EDA offers stability, EEG provides richer information about cognitive effort at the cost of higher sensitivity to noise. Preprocessing strategies thus play a decisive role, as illustrated by the improvements achieved by adapting EEG frequency bands to exclude artifact-dominated ranges. This example reinforces the broader point that understanding the characteristics of the data and the recording environment is essential before attempting classification.

Interestingly, across all signal types, the random forest classifier consistently outperformed other algorithms, suggesting that ensemble-based approaches may be particularly well-suited to handling noisy, heterogeneous data in HRC contexts.

Finally, the design of the experimental protocol warrants critical reflection. While the digit span task was effective in reliably inducing cognitive effort, its use alongside the collaborative manipulation task identifies this protocol as a double-task design. Such designs are common in HRC research because they are straightforward to implement and benefit from the robustness of validated psychological paradigms. However, they come at the cost of ecological validity and long-term usability. By shifting participants' focus toward the secondary task, the protocol risks making the HRC context peripheral rather than central, thereby limiting the insights that can be drawn about cognitive states directly elicited by collaboration with the robot. More importantly, because the robot does not itself generate the cognitive demand, its role in regulating or adapting to the participant's mental state remains minimal. This severely constrains the potential to investigate adaptive countermeasures to cognitive effort, a crucial step for future MSM-based systems. Double-task designs may be justified in scenarios that genuinely involve multitasking, such as human-robot teaming where task responsibilities are dynamically shared. Yet for most applications, it would be more appropriate to adopt protocols in which the robot itself modulates cognitive demand, ensuring that mental state changes are tightly coupled to the collaborative interaction [120].

In summary, this contribution demonstrates the feasibility of integrating highly wearable physiological sensors into HRC experiments while also exposing the significant challenges that remain. The findings stress the need for improved robustness across sessions, careful handling of robot-induced artifacts, and thoughtful protocol design that prioritizes ecological validity. Future research should build on this foundation by exploring more interactive scenarios in which the robot's behavior directly influences the participant's cognitive effort, thereby aligning experimental conditions more closely with real-world applications. Such approaches, though more complex, could elicit

richer and more generalizable insights into human mental states during collaboration with robots.

# Chapter 6

# Make it meaningful

## 6.1 Motivation

As highlighted in the previous chapter, merging Human-Robot Collaboration (HRC) with Mental State Monitoring (MSM) raises a unique set of challenges.

On the robotics side, tasks that humans execute effortlessly can still be highly challenging for robots, particularly in shared workspaces where safety requirements add to technical limitations. Beyond these constraints, realistic collaboration imposes a certain level of interdependence between both agents which requires time and freedom to be expressed. As a result, an HRC trial lasting less than a few minutes proves challenging.

On the physiological side, MSM requires large volumes of data to address the multiple layers of variability inherent in human physiology. Intra-session variability necessitates repeated trials within the same individual to robustly capture task effects. Inter-individual variability makes group-level generalization challenging, often requiring user-specific models. Inter-session variability, in turn, further limits the reusability of models trained in one session for future use with the same user. Although emerging methods offer promising solutions, the reality remains that robust MSM demands tens or hundreds of repeated and comparable trials per participant, an approach that quickly translates into long experimental sessions and significant participant burden. Ensuring comfort and engagement while avoiding confounds such as fatigue becomes critical for both ethical and methodological reasons.

Taken together, these constraints expose a fundamental tension in HRC research: collaboration takes place in long, naturalistic interactions, whereas robust physiological monitoring relies on numerous short and comparable trials. The temporal dimension of this trade-off forces researchers to bal-

ance data quantity against experimental duration and participant well-being. Equally important, however, is the methodological dimension : how to design trials that are both comparable and naturalistic. A widely adopted solution has been the use of dual-task paradigms, in which a secondary task manipulates cognitive load independently of the primary HRC scenario. This approach solves the problem by dividing it into two parts: one standardized task to induce cognitive effort and one HRC scenario to demonstrate collaboration. Yet, by inducing effort externally rather than allowing it to arise from the collaboration, dual-task designs may introduce interaction biases between the two tasks and limit the relevance of findings for real-world applications.

To address these gaps, this chapter introduces three experimental protocols designed to induce and study cognitive effort directly within HRC dynamics, while emphasizing reproducibility and scalability. Each protocol is given a label to facilitate references.

1. The first protocol (RITHMB) adapts three neuropsychological paradigms into robot-centered tasks, where the robot itself provides the stimuli. These tasks were deployed in a video-based format to ensure experimental control, comparability with the original paradigms and easier diffusion within the research community. Their evaluation relied on subjective and behavioral measures only, laying the groundwork for selecting the most suitable paradigms for future physiological studies.

2. The second protocol (RITHMD) builds on this selection to create one of the first open-access physiological databases dedicated to HRC. Two of the video-based tasks were implemented across two sessions per participant, with subjective, behavioral, and physiological measures recorded. This dataset directly addresses the sparsity of shared resources in the field and provides the means to study inter-subject, inter-session, and inter-task variability. Beyond supporting reproducibility, it provides a benchmark for developing and comparing MSM pipelines while tackling physiological variability, one of the main barriers to real-world deployment of MSM in HRC.

3. The third protocol (RITHME) complements these video-based studies by investigating the potential gap between video and physical robot conditions. One of the tasks was implemented both in video form and with a physically present robot, with subjective, behavioral, and physiological data collected. This comparison is critical: although video-based setups facilitate reproducibility and data collection, physical robot presence may elicit different cognitive processing, such as a

stronger engagement. Quantifying these differences helps clarify the extent to which findings based on video stimuli can be generalized to interactions with real robot.

Together, these three experimental campaigns tackle one of the most pressing limitations of MSM in HRC: the lack of validated paradigms and openly available datasets. By offering a set of reproducible tasks, a comprehensive database, and a direct comparison between video and physical modalities, this work provides the community with tools to better address the challenges of physiological variability and to design adaptive systems grounded in robust evidence.

To ensure transparency and facilitate broader research efforts, all necessary resources, including codes and instructional materials required for running these tasks, are openly accessible online (see Section 6.7). This commitment to open science aims to enhance reproducibility and encourage collaboration within the research community, fostering advancements in HRI research methodologies and practices.

## 6.2 Related Work

Studying cognitive effort during human-robot interaction is essential for effective collaboration [15, 177]. However research often focuses on specific use cases, leading to a variety of non-comparable results. For instance, the assembly task is popular in industrial contexts but varies widely in implementation. It is based on a human assembling an object with different levels of assistance from a cobot. Yet, the objects to assemble can be as diverse as aluminum structures [146], GPUs [55], or prototype models [178]. By inducing cognitive effort in a non-standard way, these tasks lack comparability and control over the effort induced, limiting the generalizability of the results.

For a more standardized and controlled approach, dual-task paradigms involve performing a primary ecological task alongside a secondary controlled task to induce cognitive effort. This second task can presents digits to recall while building a wall [53] or arithmetic operations to perform during a collaborative pick-and-place [179]. It can also be an auditory n-back task during robotic surgery [180] or while performing a color-matching task with a robotic arm [181]. However, dual-task paradigms come with important limitations. Interactions between the secondary and the primary task may introduce confounding effects or performance biases, thereby reducing the validity of the measured outcomes [182, 183]. Moreover, dual-task designs

do not easily translate into adaptive HRC scenarios, as they do not provide a direct pathway for adjusting the robot's behavior in real time .

Hence, new tasks should be designed to induce cognitive effort directly using the robot's behavior. Ideally, these tasks should be derived from pre-existing tasks already used in human factor research to ensure their robustness.

The N-Back task [184] is a great candidate as it is already extensively used in neuroimaging and human performance research due to its highly controlled procedure [185, 186, 187]. In this paradigm, participants are shown a sequence of stimuli and must indicate whether the current item matches the one presented n steps earlier. Thus cognitive effort is only manipulated by changing the instruction, ensuring that stimuli presentation remains consistent across different conditions[188].

The Sternberg task [189] meets an equivalent popularity [190, 191] as it allows easier disambiguation of several memory-related processes such as encoding, maintenance and retrieval. In this paradigm, participants are asked to memorize a sequence of items and then decide whether a probe item was part of that sequence. Its structured design allows researchers to systematically manipulate parameters such as sequence [192], interstimulus interval [193], and probe duration, offering fine-grained control over the cognitive demands associated with each memory subprocess.

While less established than the two others, the proposed Cognitive Shifting task is inspired by the Affective Shift task [194] which belongs to the task-switching paradigm [195, 196]. In this paradigm, participants identify the odd-one-out in a sequence based on changing rules. It is designed to induce cognitive effort in a task-switching paradigm, allowing to study potential interactions between these two effects.

## 6.3 RITHMB: Validation of an Open Access Protocol

### 6.3.1 Overview

Developing robotic tasks based on established paradigms to induce cognitive effort involved several key steps. First, several robotic movements were designed as short and comprehensive stimuli. Then, three paradigms were selected, each inducing cognitive effort in distinct ways to provide a larger perspective. Both the stimuli and paradigms were implemented to ensure ease of diffusion and reproducibility.

Finally, a user study was conducted to compare the original letter-based tasks with the robotic versions. Participants interacted with video clips of the robot instead of the physical robot to facilitate the diffusion of the tasks. In addition, this modality maintained consistency with the original stimuli, which were also presented on a screen.

### 6.3.2 Stimuli

#### Design

In order to build a quick, standardized and highly repeatable calibration scenario, short robot movements inspired from human sign langage have been designed. As pseudo-langage material, these stimuli properties are close to the ones presented by letters. Thus, they should be applicable to any experiment in which letters are used as stimuli.

Sign language structure is usually described using three components: tabula, designator, and signation [197]. The tabula of a sign refers to its spatial position, possibly in relation to the signer's body, while its designator is related to the configuration or shape of the hand(s). Its signation corresponds to the movement, or change in configuration, of the hand(s).

Because the robot chosen to perform these sign language-inspired movements had limited handshape configuration, only tabula and signation components were manipulated. In addition, it is worth noticing that these signs have been designed for research purposes and that, while inspired by sign language, they do not constitute a sign language themselves. To simplify communication and highlight this distinction, the tabula component was referred to as "localisation," and the signation component was split into "movement" and "direction".

Hence, each sign results in the robot using one arm to point at one location following a linear movement in a specific direction (see Figure 6.1). Thus, each sign can be described as a combination of four parameters:

- Laterality: left or right arm;
- Localisation: head, torso, arm, or base;
- Movement: vertical, sideway, or horizontal;
- Direction: forward or backward.

Resulting in a total of $2 * 4 * 3 * 2 = 48$ signs.

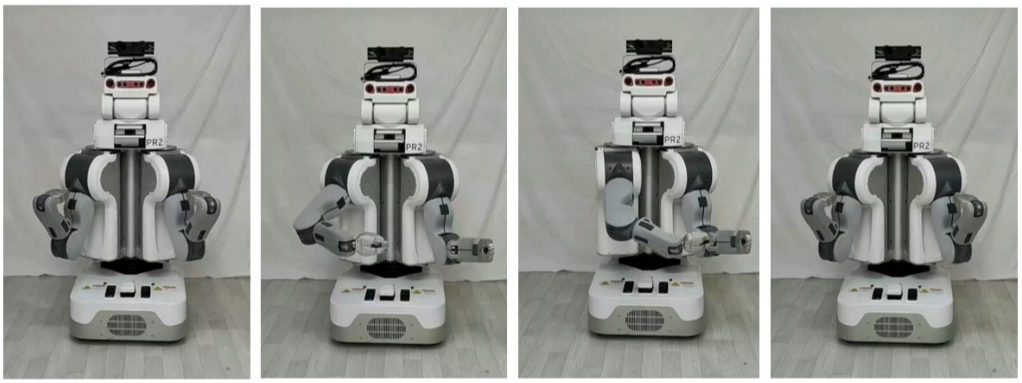

Figure 6.1: In this sign, the robot uses its **right arm** to point at its **left arm** following a **sideway** movement in a **forward** direction.

### Implementation

Stimuli have been implemented using a PR2 robot. Such a robot is composed of two arms with seven degrees of freedom, a mobile head, and a mobile base. To avoid any distraction, only the arms were controlled during a stimulus. The head remained fixed, facing the front.

Considering a common initial arms position, each sign has been implemented as a trajectory composed of four key-frames representing joint configurations (Each picture in Figure 6.1 shows a key-frame). While the first and the last ones always correspond to the initial position, the two others correspond to the sign itself for the signing arm. In a forward direction, the second keyframe defines the approach position and the third keyframe the pointing position. During a stimulus, the second arm is kept at its initial position unless the location is the arm. In this latter case, the arm is deployed on the side at a fixed position to create readable signs.

The signs were first implemented using a digital twin of the robot and the 3D software Blender. This method is of great interest in an interdisciplinary team as manipulating the digital twin does not require any programming knowledge. To reduce the generation complexity and to ensure the homogeneity over the signs, these later have been only designed using a single laterality and the backward direction. The second direction, namely forward, has then been automatically generated by inverting the two first key-frames for each sign and the opposite laterality has been generated by taking the invert of the shoulder pan joint, elbow flex joint, and wrist flex joint.

To ensure short movements while avoiding overshooting effects due to too fast movements, the interpolation between two arm configurations has

been settled to 1.05 second providing 3.15-second long signs. Fine-tuning has been realized on a real Pr2 robot to validate signs readability, joint limits compliance, and correct some remaining joint overshoots. The link to the script resulting from this fine-tuning is available at Section 6.7

Finally, the same Pr2 robot has been filmed while performing all signs. This sequence has been edited into videoclips of 3.7 seconds, which have been used as robot stimuli for all the tasks described in this paper. Although this duration is longer than usual letter or number presentation in working memory tasks, it was chosen to keep movements smooth.

### 6.3.3 Tasks

#### Design

To systematically investigate cognitive effort within HRC, three experimental tasks were adapted from well-established paradigms in cognitive psychology and neuroscience: the N-back task, the Sternberg task and the Cognitive Shifting task (see Section 6.2). Each task was carefully designed to preserve the core cognitive mechanisms described in the literature.

In the N-back task, participants are presented with a sequence of stimuli (see Figure 6.2). For each stimulus, they must indicate whether the current stimuli matches the one shown n trials ago. A higher n is expected to induce a higher cognitive effort.

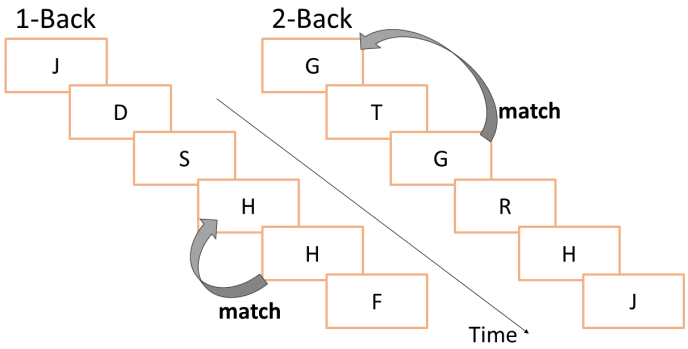


Figure 6.2: N-Back task experimental design with letters as stimuli.

In the Sternberg task, for each trial, participants are presented with a sequence of stimuli (see Figure 6.3). After a brief delay, they are shown a single probe and asked to indicate whether the probe was present in the initial

set. A longer sequence of stimuli is expected to induce a higher cognitive effort.

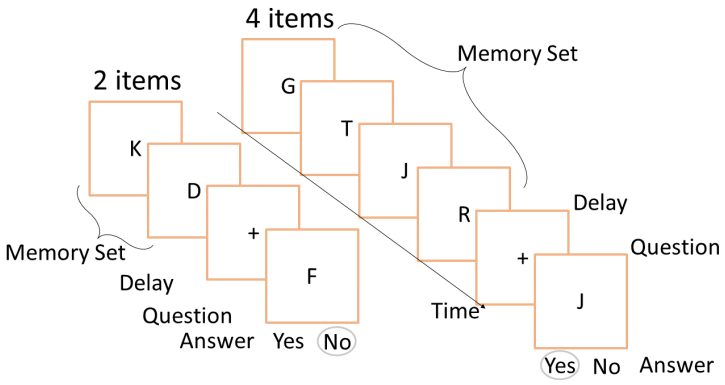


Figure 6.3: Sternberg task experimental design with letters as stimuli.

In the Cognitive Shifting task, for each trial, participants are presented with a sequence of four stimuli (see Figure 6.4). Stimuli can vary over three parameters for letters (position, color, rotation) as for robot video clip (localisation, laterality, movement) and among the four stimuli, three can be considered as odd ones (one per parameter). Participants need to perform an odd-one-out search based on the relevant(s) parameter(s) indicated by the cue. At the end of the trial, participants are asked to select the odd-one-out based on one parameter among the ones presented in the cue. The more parameters are presented in the cue, the higher cognitive effort is expected to rise.

### Implementation

The three tasks were implemented in python using psychopy 2022.2.5 [198]. The repository (see Section 6.7) contains for each task the .psyexp file editable with psychopy builder along with the _lastrun.py file which can be run directly with python. In addition, two python jupyter notebooks allow to generate an appropriate csv file for task configuration and result analysis respectively.

For this specific study, all tasks were configured to be performed during 6 blocks, alternating between robot and letter stimuli from one block to the other. Each block presented the 3 conditions in a pseudo-random order but conditions were not mixed together: all trials from one condition had to be performed before moving on to the next one (see Figure 6.5).

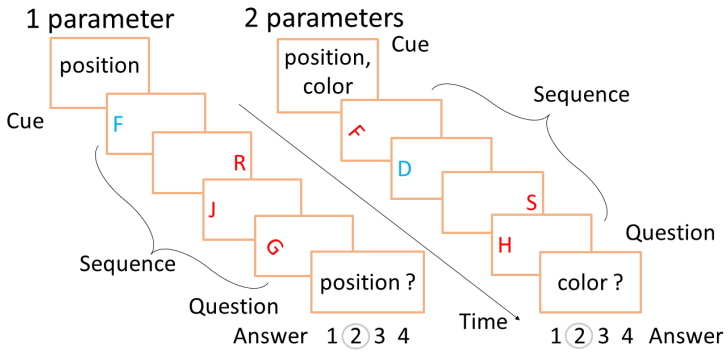


Figure 6.4: Cognitive Shifting task experimental design with letters as stimuli.

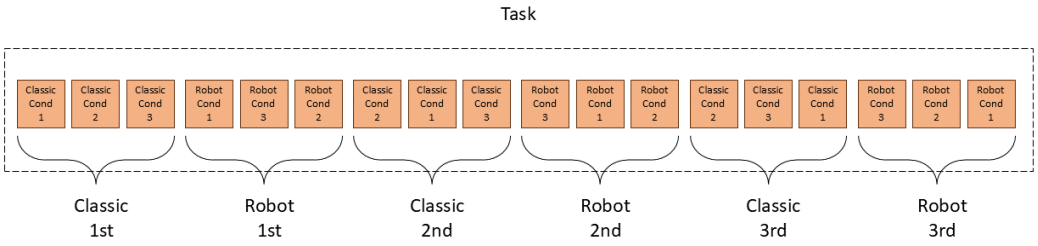


Figure 6.5: Experimental protocol for the RITHMB study: 6 blocks with a block presenting either only the letter version or only the robot version of the task. Each block presents the 3 conditions in a pseudo-random order.

The robot stimuli were the above-mentioned ones and lasted around 3.7 seconds. The letter stimuli were white and lasted 2 seconds. Given these durations and other task-dependent parameters (see Table 6.1, the global duration of each task was kept around an hour ).

In the N-Back task, the three conditions corresponded to the values of N : 1, 2 and 3. A condition contained 50 trials per block (150 in total) with 20% of targets. It resulted in 260 seconds to perform a full block of 150 trials with robot stimuli, while it required 180 seconds with letter stimuli.

For the Sternberg task, the three conditions corresponded to the sequence's size : 2, 4 and 6 items. A condition contained 10 trials (30 in total) with half of the presented sequences containing the probe. It resulted in 225 seconds to perform a full block of 30 trials with robot stimuli, while it required 135 seconds with letter stimuli.

For the Cognitive Shifting task, the three conditions corresponded to the

number of parameters to monitor : 1, 2 and 3 parameters. A condition contained 9 trials (27 in total) with each parameter being used 3 times as criteria for odd-one selection. It resulted in 200 seconds to perform a full block of 27 trials with robot stimuli, while it required 135 seconds with letter stimuli.

| **Task** | **N-Back** | **Sternberg** | **Cog Shifting** |
|---|---|---|---|
| **Stimuli Type** | Robot / Letter | Robot / Letter | Robot / Letter |
| **Total Blocks** | 3 / 3 | 3 / 3 | 3 / 3 |
| **Total Trials** | 450 / 450 | 90 / 90 | 81 / 81 |
| **Block Duration (s)** | 260/180 | 225/135 | 200/135 |
| **Block parameters** | | | |
| **Number of conditions** | 3 | 3 | 3 |
| **Trials per Condition** | 50 | 10 | 9 |
| **Target Percentage** | 20% | 50% | N/A |

Table 6.1: Summary table of experimental design parameters in the RITHMB study for the N-Back, Sternberg, and Cognitive Shifting tasks.

### Measures

The same subjective and behavioral metrics are collected during each task. Cognitive effort is subjectively reported by participants using the Rating Scale Mental Effort (RSME) which appears at the end of each condition [199]. A particular attention was given to the position of the labels [200].

The key input of each trial allows to collect response time along with an answer which is subsequently used to compute the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) for each condition. While response time directly gives a value in second, other metrics need to be aggregated in an accuracy score. Given the interdisciplinary nature of the HRI field, two formulas have been used to compute the accuracy score.

A first formula computes the percentage of correct answers as usually done in psychology and cognitive science (see Equation 6.1). This metric gives an intuitive, global measure of correctness by considering all correct responses, whether to target or non-target stimuli.

$$\%correct = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

A second formula computes the F1 score widely used in machine-learning as the harmonic mean between precision and recall. This metric allows to model two slightly different measures of performance. First, the precision refers to the number of correct answers among all the trials identified as target (see Equation 6.2). Second, the recall models the proportion of recognized targets among all the trials that indeed contained a target (see Equation 6.3).

$$Precision = \frac{TP}{TP + FP} \tag{6.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{6.3}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6.4}$$

Further studies could also attempt to model performance more precisely by balancing precision and recall differently depending of their end goal. Yet, the F1 score was kept here for its simplicity and popularity (see Equation 6.4).

Hence, by reporting both the percentage of correct answers and the F1 score, this study provides complementary insights suited to a broad scientific audience. This dual approach enhances interpretability across disciplines and enables a more nuanced analysis of performance variations across the tasks.

### 6.3.4 User studies

#### Recruitment

For each task, 24 participants were recruited via advertisements disseminated on the campuses of two universities, for a total of 72 participants. This number was assessed through power calculation using G*Power software and effect size reported in studies with similar task parameters[186, 193]. The number or participants, their age, gender, education, professional situation and prior knowledge are reported for each task in table 6.2.

#### Hypotheses

To validate the cognitive effort induction an effect of difficulty levels on performance was expected. More precisely, as task difficulty increased:

- Accuracy should decrease;
- Subjective effort and response time should increase.

| | N-back | Sternberg | Cog Shifting |
|---|---|---|---|
| **N** | 23 | 21 | 22 |
| **Age** | 27.91 ± 5.51 | 26.43 ± 6.18 | 26.36 ± 6.16 |
| **Gender** | Male: 15<br>Female: 8 | Male: 14<br>Female: 7 | Male: 15<br>Female: 7 |
| **Highest Education Level** | Graduate: 14<br>Undergraduate: 4<br>Doctorate: 3<br>Highschool: 2 | Graduate: 9<br>Undergraduate: 9<br>Doctorate: 1<br>Highschool: 1 | Graduate: 11<br>Undergraduate: 7<br>Doctorate): 4 |
| **Professional Situation** | Student: 11<br>PhD Student: 3<br>Employed: 7<br>Unemployed: 2 | Student: 11<br>PhD Student: 6<br>Employed: 3<br>Unemployed: 1 | Student: 10<br>PhD Student: 7<br>Employed: 5 |
| **Knowledge in** | | | |
| **Robotics** | 2.3 ± 1.02 | 2.43 ± 1.08 | 2.64 ± 0.9 |
| **Physio** | 2.3 ± 1.18 | 2.71 ± 1.31 | 3.05 ± 1.33 |
| **Sign** | 1.26 ± 0.45 | 1.24 ± 0.54 | 1.59 ± 0.91 |

Table 6.2: Participant demographics and knowledge metrics in the RITHMB study for N-back, Sternberg, and Cognitive Shifting tasks.

In addition to this usual effect of task difficulty, a stimuli effect was also expected to affect results. The robot condition was expected to be harder than the letter condition. Indeed, robot stimuli deliver much more complex visual information. This would result, at the metric level, in the same variations as an increase in difficulty, i.e. decreased accuracy, increased subjective effort and increased response time.

### Protocol

Once recruited, participants booked a time slot with the experimenter via email. Upon arrival, they were greeted and asked to sign a consent form, detailing the study's purpose, procedures, risks, and benefits. After checking for inclusion and exclusion criteria, they completed a demographic questionnaire.

The experimenter explained the task and conducted a brief training session to familiarize participants with it. Once participants understood the instructions and had no further questions, the experiment began.

The whole experiment was performed as described in 6.3.3 and only required a Dell Precision T1700 desktop computer using Windows 10 operating system, with no specialized hardware except a keyboard and a mouse.

After completing the tasks, participants were debriefed. The experimenter provided an in-depth explanation of the study's goals, answered any questions, and distributed monetary compensation as a token of appreciation for their time. The whole procedure lasted between 1 and 1.5 hour depending on the task.

This protocol has received approval from the institutional ethics review board of the University of Toulouse (project n°2024_839).

#### Analyses

Subjects were considered as outliers if one of their behavioral results (accuracy, response time) diverged from others of more than 3 standard deviation in at least one condition. This rule led to the rejection during results analysis of 1 subject for the N-back task, 3 subjects for the Sternberg task and 2 subjects for the Cognitive Shifting task. Yet, full data are still available in the shared repository.

Data analysis has been performed using python, with the pingouin library for statistics [201]. Given the experimental designs, a two-ways RMANOVA was applied to each metric of each task for stimuli effect, difficulty level effect and interactions between them. In case of a significant effect, and after testing for normality, pairwise t-tests were then performed with Bonferroni correction.

### 6.3.5 Statistical Results

#### N-Back

The statistical results for the N-back task revealed significant effects of difficulty level, stimuli, and of their interaction across all metrics. A synthetic view of these results can be found in Figure 6.6.

The effect of difficulty level was significant across all metrics: % correct ($F(2, 44) = 56.12, p < .001, \eta_p^2 = .72$) , F1 score ($F(2, 44) = 148.37, p < .001, \eta_p^2 = .87$), Response Time ($F(2, 44) = 68.12, p < .001, \eta_p^2 = .76$), and RSME ($F(2, 44) = 87.95, p < .001, \eta_p^2 = .80$). For all metrics, posthoc tests revealed significant differences between all pairs of difficulty level, with a bigger size effect for the difference between the first and last difficulty level than for other pairs.

The effect of stimuli was also significant across all metrics: % correct

($F(1, 22) = 33.22$, $p < .001$, $\eta_p^2 = .60$) , F1 score ($F(1, 22) = 48.28$, $p < .001$, $\eta_p^2 = .69$), Response Time ($F(1, 22) = 660.77$, $p < .001$, $\eta_p^2 = .97$), and RSME ($F(1, 22) = 17.97$, $p < .001$, $\eta_p^2 = .45$).

The interaction between stimuli and condition was only significant for the F1 score ($F(2, 44) = 6.71$, $p < .01$, $\eta_p^2 = .23$). For this metric, posthoc tests revealed significant differences between the first difficulty level and the two others for the letter stimuli, while for the robot condition stimuli differences were significant between all pairs of difficulty levels.

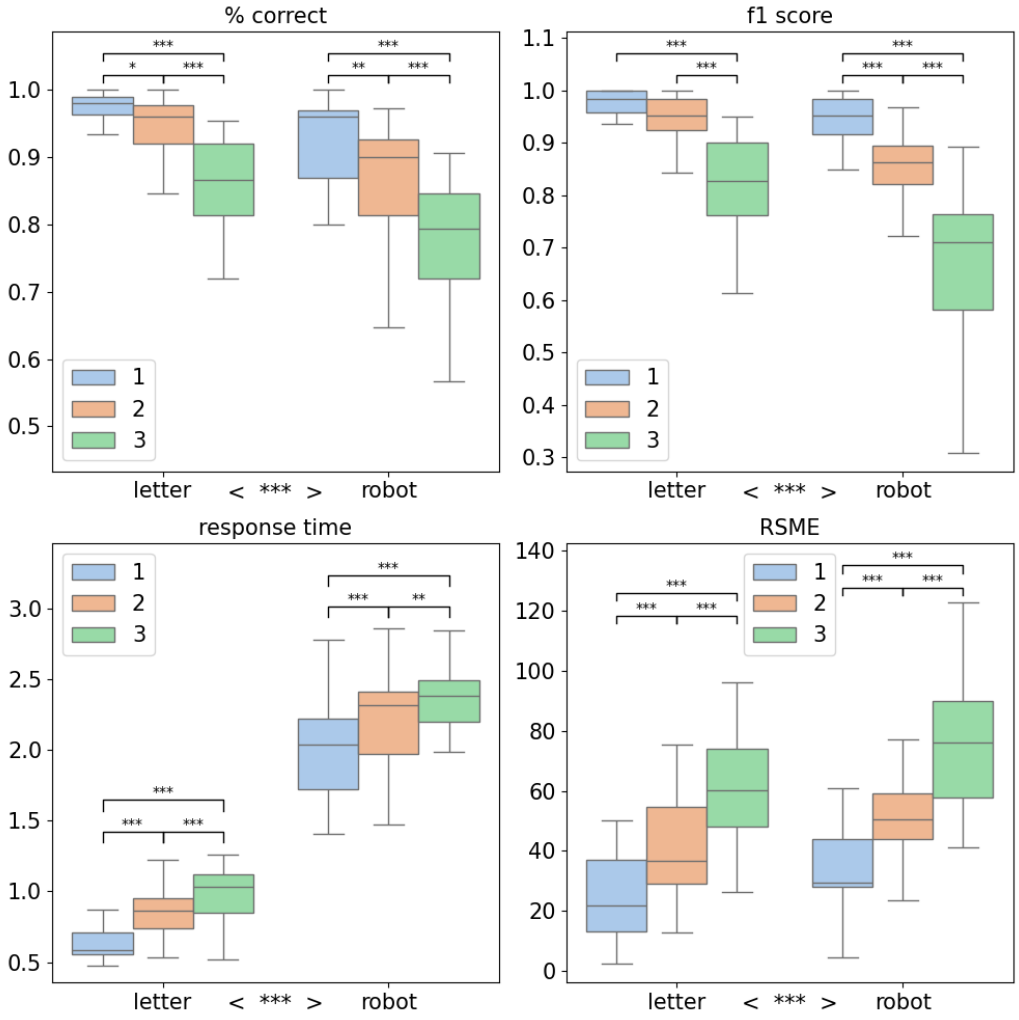


Figure 6.6: Results for the N-Back task over the four metrics. The Stimuli effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

### Sternberg

The statistical results for the Sternberg task revealed significant effects of difficulty level, stimuli, and of their interaction across all metrics. A synthetic view of these results can be found in Figure 6.7.

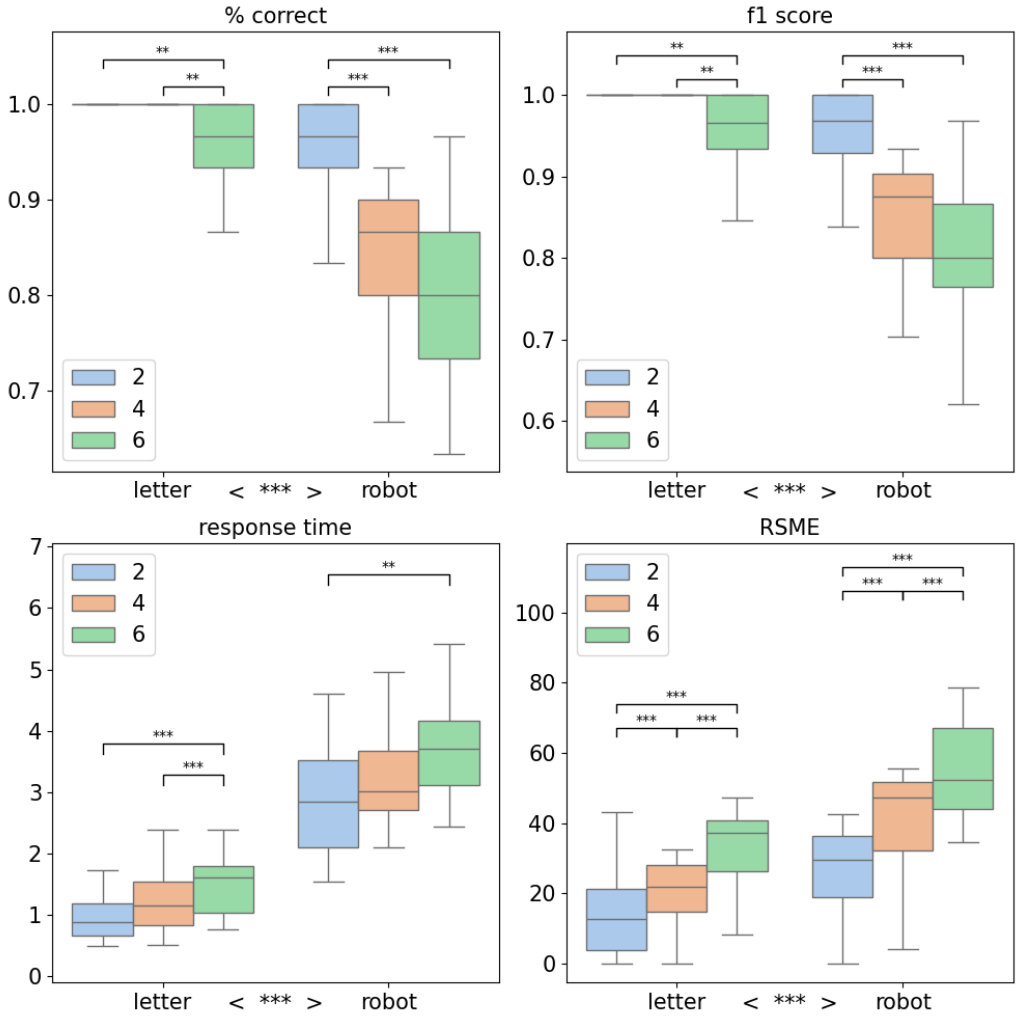


Figure 6.7: Results for the Sternberg task over the four metrics. The Stimuli effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

The effect of difficulty level was significant across all metrics: % correct ($F(2, 40) = 37.13$, $p < .001$, $\eta_p^2 = .65$), F1 score ($F(2, 40) = 33.11$, $p < .001$, $\eta_p^2 = .62$), Response Time ($F(2, 40) = 17.16$, $p < .001$, $\eta_p^2 = .46$), RSME ($F(2, 40) = 102.02$, $p < .001$, $\eta_p^2 = .84$). For all metrics except response time, posthoc tests revealed significant differences between all pairs of difficulty levels, with a bigger size effect for the difference between the first and last

difficulty levels than for other pairs. When considering response time, only differences between the last difficulty level and the two others were significant.

The effect of stimuli was also significant across all metrics: % correct ($F(1, 20) = 71.0$, $p < .001$, $\eta_p^2 = .78$) , F1 score ($F(1, 20) = 72.95$, $p < .001$, $\eta_p^2 = .78$), Response Time ($F(1, 20) = 113.18$, $p < .001$, $\eta_p^2 = .85$), and RSME ($F(1, 20) = 71.05$, $p < .001$, $\eta_p^2 = .78$).

The interaction between stimuli and condition was only significant for the metrics related to accuracy, namely % correct ($F(2, 40) = 14.97$, $p < .001$, $\eta_p^2 = .43$) and F1 score ($F(2, 40) = 13.37$, $p < .001$, $\eta_p^2 = .40$). For these metrics, posthoc tests revealed significant differences between the last condition and the two others for letter stimuli while for robot stimuli differences were significant between the first condition and the two others.

### Cognitive Shifting

The statistical results for the Cognitive Shifting task revealed significant effects of difficulty level, stimuli, and of their interaction across all metrics. A synthetic view of these results can be found in Figure 6.8.

The effect of difficulty level was significant across all metrics: % correct ($F(2, 42) = 43.98$, $p < .001$, $\eta_p^2 = .68$), F1 score ($F(2, 42) = 39.89$, $p < .001$, $\eta_p^2 = .66$), Response Time ($F(2, 42) = 107.41$, $p < .001$, $\eta_p^2 = .84$), and RSME ($F(2, 42) = 45.68$, $p < .001$, $\eta_p^2 = .69$). For all metrics posthoc tests revealed significant differences between all pairs of difficulty levels, with a bigger size effect for the difference between the first and last difficulty levels than for other pairs.

The effect of stimuli was also significant across all metrics: % correct ($F(1, 21) = 42.98$, $p < .001$, $\eta_p^2 = .67$) , F1 score ($F(1, 21) = 41.16$, $p < .001$, $\eta_p^2 = .66$), Response Time ($F(1, 21) = 29.72$, $p < .001$, $\eta_p^2 = .59$, and RSME ($F(1, 21) = 81.26$, $p < .001$, $\eta_p^2 = .79$).

The interaction between stimuli and condition was only significant for the metrics related to accuracy, namely % correct ($F(2, 42) = 16.75$, $p < .001$, $\eta_p^2 = .44$) and F1 score ($F(2, 42) = 13.94$, $p < .001$, $\eta_p^2 = .40$). For these metrics, posthoc tests revealed significant differences between the first condition and the two others for letter stimuli, while for robot stimuli differences were significant between all pairs.

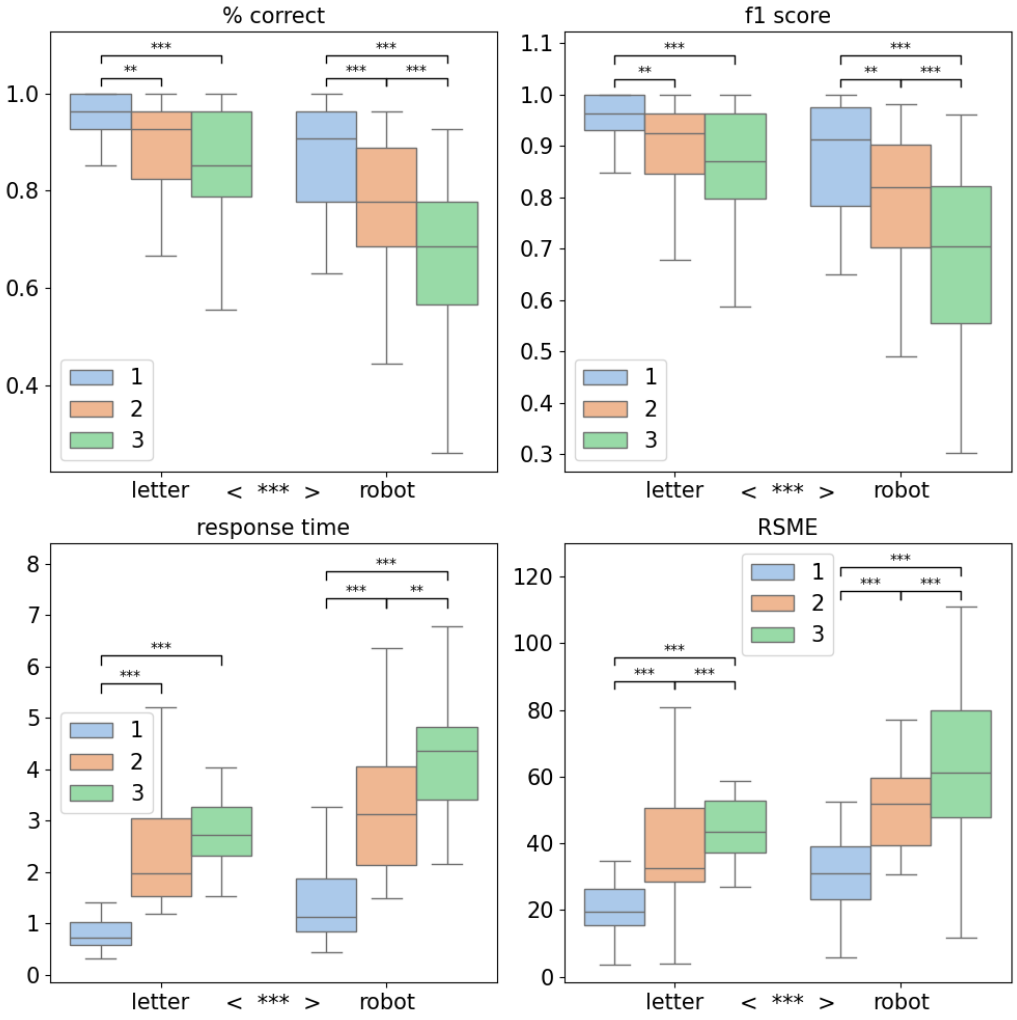


Figure 6.8: Results for the Cognitive Shifting task over the four metrics. The Stimuli effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

### Additional remarks

As a basis for further utilization, it may be noted that the N-Back task was the least appreciated, most likely due to its highly repetitive design. In addition, in the robot condition of the Cognitive Shifting task, participants found it harder to perform the odd-one-out task with the movement parameter than

with the localisation or laterality parameters.

### 6.3.6 Main outcomes

In summary, this first campaign confirmed that the robotic implementations of the N-Back, Sternberg, and Cognitive Shifting tasks effectively elicited cognitive effort, as reflected in accuracy, response times, and subjective ratings. While replicating classical effects, robotic stimuli consistently imposed additional cognitive demand compared to letter-based versions. Differences between tasks also emerged: the N-Back scaled reliably but suffered from low popularity among participants, the Sternberg task revealed steeper difficulty increases in its robotic form than with its letter form, and the Cognitive Shifting task provided a novel yet robust framework for probing effort. Together, these findings establish robotic paradigms as tools for investigating cognitive effort in collaborative contexts.

## 6.4 RITHMD: a Physiological Database

### 6.4.1 Overview

In the previous experimental campaign, cognitive tasks originating from experimental psychology and neuroscience were adapted to a robotic context through the use of short robot video clips. In this setup, the robot's behavior alone was used to induce different levels of cognitive effort, which were assessed through subjective ratings and behavioral metrics. This first campaign of studies emphasized ease of testing and reproducibility, deliberately avoiding physiological recordings that would have introduced heavy constraints in terms of setup time, invasiveness, and data handling. By relying only on video-based stimuli, the protocol required nothing more than a standard computer to be reproduced, enabling the evaluation of three different tasks at three levels of difficulty using two types of stimuli (letters vs. robot).

The outcomes of this first campaign provided valuable insights into how cognitive effort during Human-Robot Collaboration could be manipulated using only robot behavior. Building on this knowledge, the current contribution focuses specifically on the Sternberg task and the Cognitive Shifting task to collect physiological data across two sessions, with two tasks per session, hence providing to the community resources to tackle the physiological variability problem across subjects, sessions, and tasks.

### 6.4.2 Method

#### Tasks

For both the Sternberg task and the Cognitive Shifting task, robot videos were again preferred over live robotic interaction, as they ensure identical trials across participants and avoid the risk of robot-related artifacts. Since the resulting dataset is intended for broad sharing within the community, data quality and standardization were here prioritized.

The Sternberg robotic task used pseudo-signs performed by a robot as main stimuli (Figure 6.9). In each trial, participants were presented with a sequence of either two or six videos, followed by a brief delay and then a probe video. They were instructed to indicate whether the probe was part of the initial sequence. Increasing the sequence length from two to six items was expected to induce higher memory load and, consequently, greater cognitive effort. This implementation follows the structure used in our previous study but restricts the design to two levels of difficulty to increase contrast between conditions.

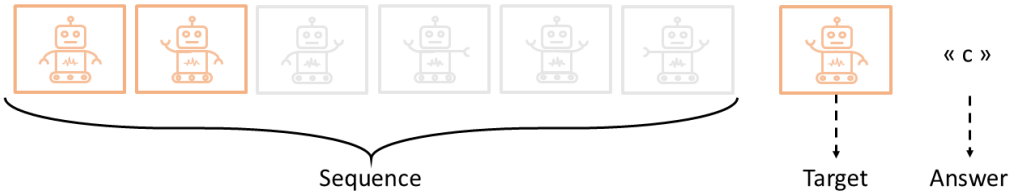


Figure 6.9: Example of Sternberg robotic task trial. For explanatory purposes, the easiest condition is colored, while the elements added for the harder condition are in grey.

The Cognitive Shifting robotic task (Figure 6.10) was implemented with either one or two pre-trial cues. For each trial, participants were presented with a sequence of four pseudo-sign varying along three possible parameters: localization (head, torso, arm, base), laterality (left vs. right), and movement type (horizontal, vertical, toward participant). In each sequence, three signs shared the same value for a given parameter, while one differed (e.g., three left-arm movements and one right-arm movement). Thus, depending on which parameter was cued, multiple signs could potentially be considered the "odd one out." To guide participants, a cue specifying the relevant parameter(s) was presented before the sequence. Single-parameter cues are

expected to create low difficulty, as they reveal directly the targeted parameter, while two-parameter cues increase ambiguity and cognitive demand. As with the Sternberg task, this design maintains continuity with earlier study. Yet, the highest difficulty level was dropped to maintain a consistent level of difficulty between tasks and avoid disengagement.

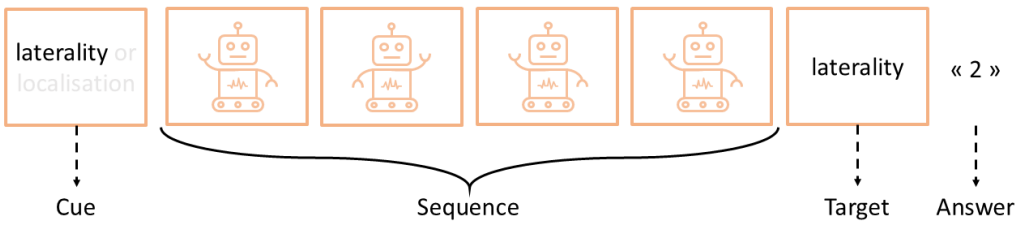


Figure 6.10: Example of Cognitive Shifting robotic task trial. For explanatory purposes, the easiest condition is colored, while the elements added for the harder condition are in grey.

### Measures

As in the previous study, the Rating Scale Mental Effort (RSME) was employed to capture subjective perception of cognitive effort [199]. This scale is particularly well-suited for experimental contexts as its simplicity allows subjective measures to be collected repeatedly across multiple time points, thus providing a finer temporal resolution of perceived effort. A particular attention was given to the position of the labels [200].

The key input of each trial allows to collect response time along with an answer which is subsequently used to compute the number of true positive, true negative, false positive, and false negative for each condition. While response time (RT) directly gives a value in second, other metrics need to be aggregated in an accuracy score. As in the previous contribution, two formulas have been used to compute the accuracy score, resulting in two slightly different metrics: % correct and F1 score.

### Physiological Sensors

In order to collect physiological activity with a suitable degree of precision, signal quality was preferred over wearability for this study (see Figure 6.11). Thus, brain activity was collected through electroencephalography (EEG)

using a 32 channels actiCHamp (Brain Products) with wet electrodes placed on a cap according to the 10/20 international system and sampled at 500 Hz. In addition, cardiac activity sampled at 1000 Hz was collected through electrocardiography (ECG) using a Faros 360°. It was also collected through photoplethysmography (PPG), along with electrodermal activity (EDA), using a Shimmer3 GSR+ Unit sampled at 128 Hz.

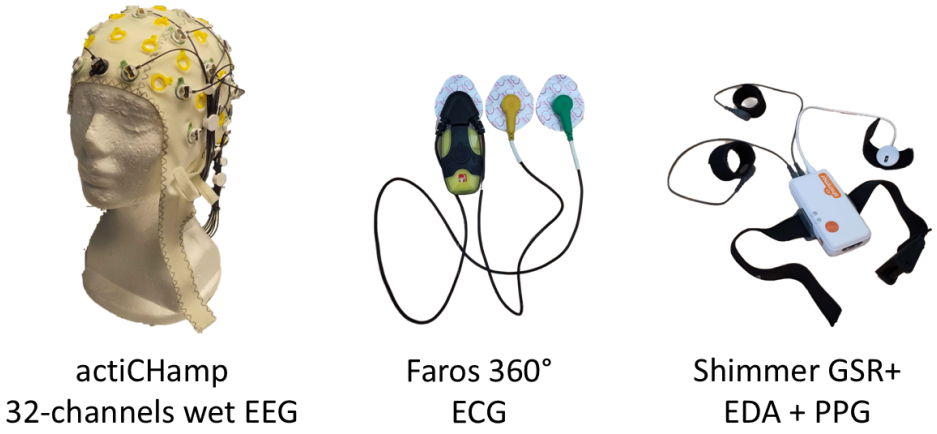


Figure 6.11: Illustration of the sensors employed in the RITHMD study for physiological data acquisition : an actiCHamp system with 32-channel, a Faros 360° device and a Shimmer3 GSR+ Unit.

### 6.4.3 User Study

#### Recruitment

30 participants were recruited via advertisements disseminated on the campuses of two universities. Their age, gender, education and professional situation are reported in table 6.3. In addition, participants' knowledge in robotics, physiology and sign language as also been collected on a 5-points likert scale.

#### Protocol

Each participant completed two experimental sessions. The overall structure of the sessions was identical, ensuring that participants encountered the same

| Gender | Highest Education | Professional Status | Knowledge |
|---|---|---|---|
| Female: 15 | Graduate: 17 | Student: 13 | Robotics: 2.3 ± 0.8 |
| Male: 15 | Undergraduate: 7 | PhD Student: 11 | Physio: 3.0 ± 1.4 |
| | Doctorate: 4 | Employed: 5 | Sign: 1.3 ± 0.5 |
| | Highschool: 2 | Unemployed: 1 | |

Table 6.3: Participant demographics and knowledge metrics for the 30 participants. Mean age was equal to 25.0 ± 4.3.

paradigms and conditions in the same order across sessions. The only differences were procedural: session 1 began with the completion of an informed consent form and a demographic questionnaire, while session 2 concluded with a debriefing in which the experimenter explained the study's objectives in greater depth and answered participants' questions.

At the beginning of each session, participants were welcomed and completed a short training phase consisting of three practice trials for each condition of both tasks. Participants were encouraged to ask any question that arises, and the experimenter offered the option to repeat the training. However, no participant requested additional practice. Following training, the physiological sensors were installed. Wet EEG electrodes were positioned on the scalp, and EDA/PPG sensors were attached to the fingers by the experimenter. For privacy and comfort reasons, the ECG electrodes on the torso were placed by the participants themselves, following a schematic guide.

The main task then followed, lasting approximately one hour. It consisted of 12 blocks of 12 trials each. After each block, participants completed the Rating Scale for Mental Effort (RSME) and had the opportunity to take a break. Block order was structured so that participants always completed the two conditions of a given task type in randomized order before moving to the next. Thus, task type alternated every two blocks. To balance order effects, half of the participants began with the Sternberg task, while the other half began with the Cognitive Shifting task.

At the end of the session, sensors were removed and participants received monetary compensation for the session.

This protocol has received approval from the institutional ethics review board of the University of Toulouse, in the same project as the previous campaign (project n°2024_839).

### Analyses

Subjects were considered as outliers if one of their behavioral results (accuracy, response time) diverged from others of more than 3 standard deviation in at least one condition. This rule led to the rejection during results analysis of 3 subjects. In addition, due to sensor issues, a supplementary participant was excluded. While no data will be excluded from the database, excluding these outliers from the analyses was necessary to first assess the legitimity of the protocol.

Data analysis has been performed using python, with the pingouin library for statistics [201]. Given the experimental designs, a two-ways RMANOVA was applied to each metric for session effect, difficulty level effect and interactions between them. In case of a significant effect, and after testing for normality, pairwised t-tests were then performed with Bonferonni correction.

For physiological signals, preprocessing and classification followed the same procedures as detailed in the previous chapter (see Section 5.4.3 and Section 5.6.2). The only adaptation concerned the choice of temporal windows, which was task-dependent. In the present case, the size of the temporal window is limited by the shortest trials available. This corresponds to the easy condition of the Sternberg task, consisting of only two robot videos. Accordingly, ECG, PPG, and EDA signals were segmented into epochs spanning the two videos preceding the participant's response for each condition of each paradigm, yielding a temporal window of 7 seconds. Because of the faster dynamics of brain activity, EEG signals were instead epoched for each of these two videos separately, resulting in twice as many epochs as for the other signals but with a duration of 3.5 seconds each. Although using all available videos were initially considered, the final choice favored this configuration to preserve a balanced and comparable design across paradigms and difficulty levels.

## 6.4.4 Statistical Results

The results are reported separately for each task and the impact of each task on cognitive effort will be studied at the classification level.

### Sternberg

In the Sternberg paradigm, for RSME score, the effect of difficulty was significant ($F(1, 25) = 164.06$, $p < .001$, $\eta_p^2 = .86$). In addition, the interaction between session and difficulty reached significance ($F(1, 25) = 5.68$, $p < .05$, $\eta_p^2 = .18$). Posthoc tests showed that effort ratings were consistently higher in

the hard condition compared to the easy condition, in both sessions (session 1: $p < .001$, $\eta^2 = .52$; session 2: $p < .001$, $\eta^2 = .47$).

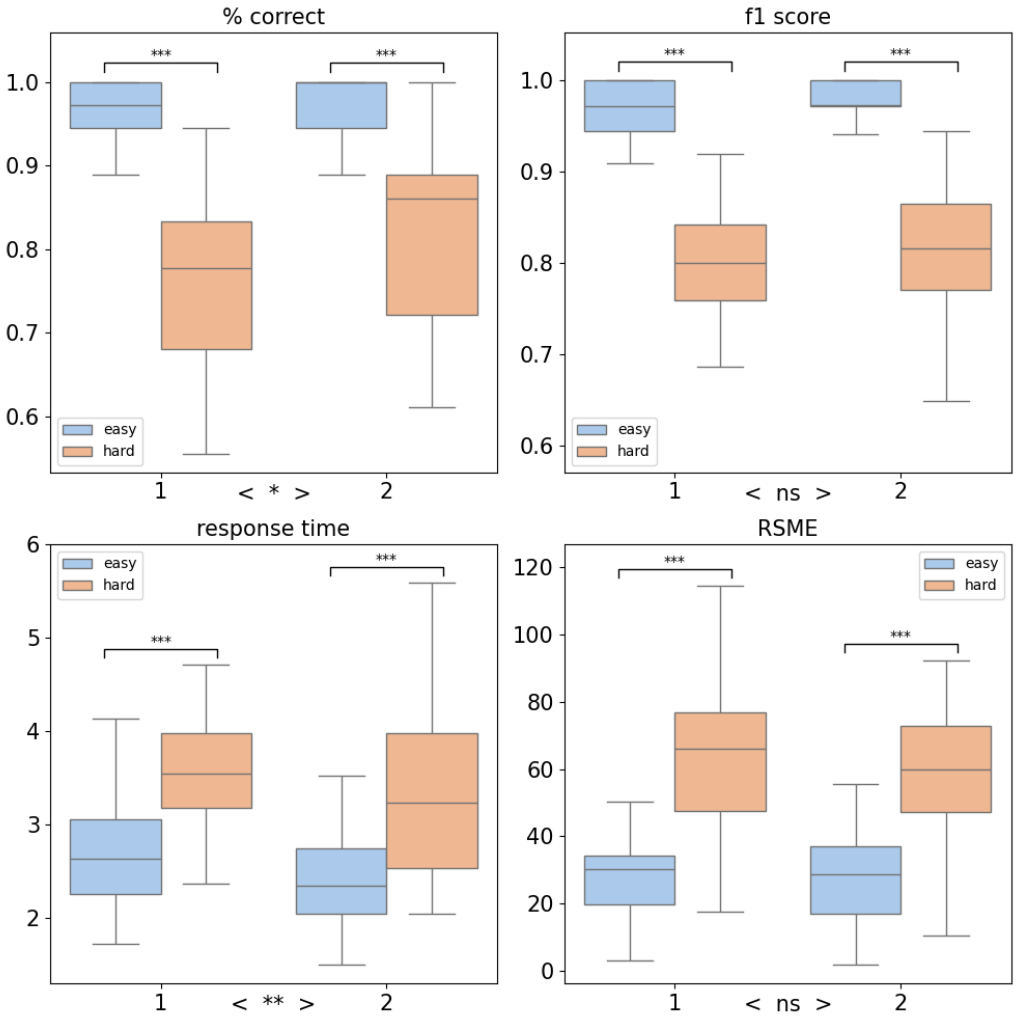


Figure 6.12: Results for the Sternberg task over the four non-physiological metrics. The session effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

For % correct, a significant effect of session was found ($F(1, 25) = 5.99$, $p < .05$, $\eta_p^2 = .19$), with more correct answer in session 2 than in session 1. An effect of difficulty was also validated ($F(1, 25) = 101.16$, $p < .001$, $\eta_p^2 = .80$).

Posthoc comparisons confirmed that % correct decreased significantly from easy to hard conditions in both sessions (session 1: $p < .001$, $\eta^2 = .58$; session 2: $p < .001$, $\eta^2 = .46$).

For the F1 score, similarly, a significant effect of difficulty was observed ($F(1, 25) = 242.04$, $p < .001$, $\eta_p^2 = .90$). Posthoc analyses revealed a performance drop between easy and hard conditions in both sessions (session 1: $p < .001$, $\eta^2 = .63$; session 2: $p < .001$, $\eta^2 = .63$).

For RT, session effect was significant ($F(1, 25) = 10.79$, $p < .01$, $\eta_p^2 = .29$), with higher values in session 2 than in session 1. In addition, a difficulty effect was found ($F(1, 25) = 133.81$, $p < .001$, $\eta_p^2 = .84$). Posthoc comparisons confirmed slower response times in hard compared to easy conditions across both sessions (session 1: $p < .001$, $\eta^2 = .32$; session 2: $p < .001$, $\eta^2 = .28$).

For EEG, difficulty effects emerged only at central site, both for beta power and the engagement index. For central beta, the effect of difficulty was significant ($F(1, 25) = 10.96$, $p < .01$, $\eta_p^2 = .30$), with post-hoc comparisons confirming higher beta activity in the hard compared to the easy condition (session 1: $p = .011$, $\eta^2 = .01$). For the central engagement index, difficulty also had a significant effect ($F(1, 25) = 5.24$, $p < .05$, $\eta_p^2 = .17$), although no post-hoc contrasts reached significance.

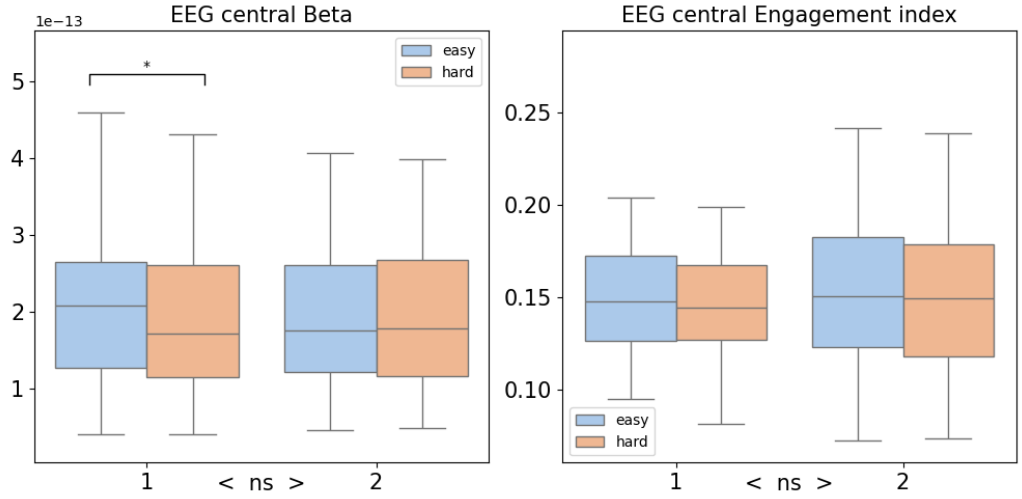


Figure 6.13: Results for the Sternberg task over EEG metrics. The session effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

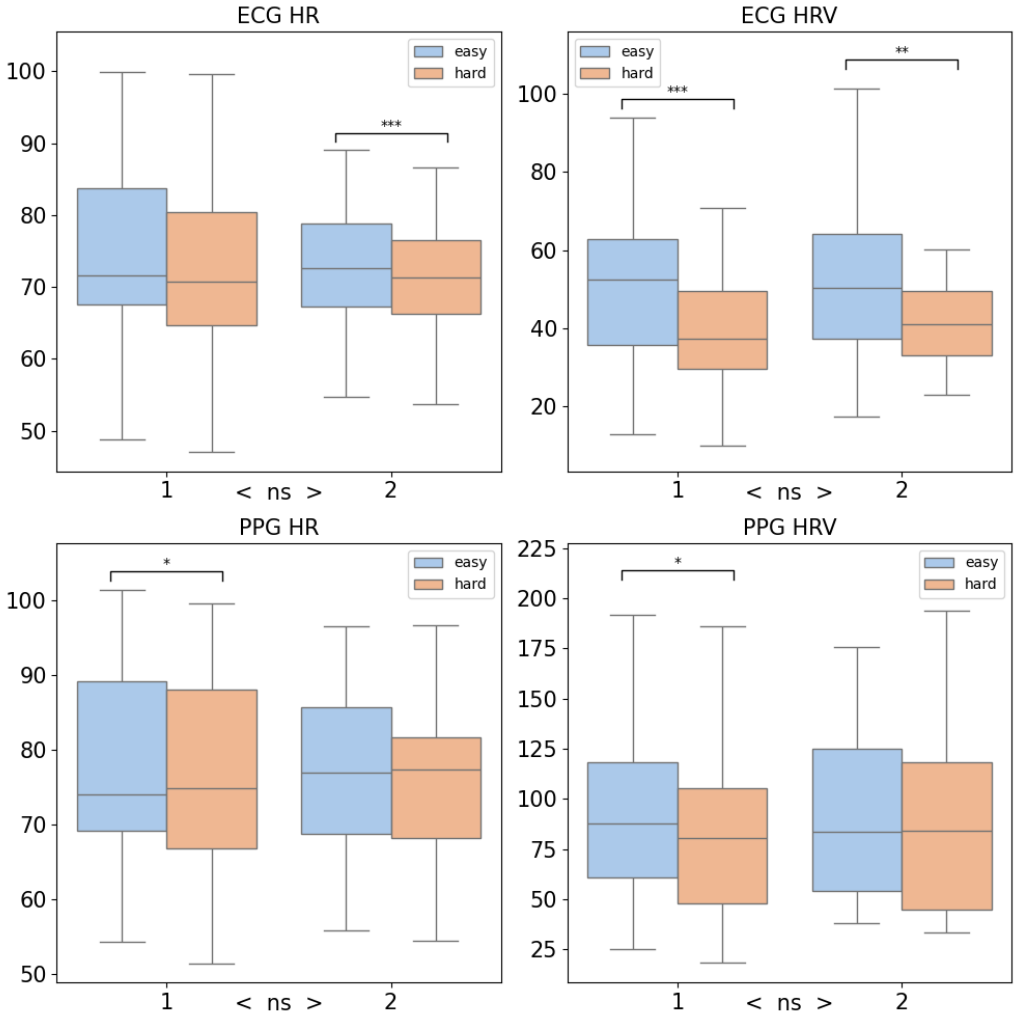


Figure 6.14: Results for the Sternberg task over ECG and PPG metrics. The session effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

For ECG, mean heart rate decreased significantly with difficulty ($F(1, 25) = 22.7$, $p < .001$, $\eta_p^2 = .48$), with an additional interaction between session and difficulty ($F(1, 25) = 4.49$, $p < .05$, $\eta_p^2 = .15$). Post-hoc tests confirmed reduced heart rates in the hard compared to the easy condition (session 2: $p = .0001$, $\eta^2 = .01$). Similarly, HRV decreased with difficulty ($F(1, 25) = 23.99$, $p < .001$, $\eta_p^2 = .49$), with post-hoc analyses revealing ro-

bust differences across both sessions (session 1: $p < .0001$, $\eta^2 = .05$; session 2: $p = .0028$, $\eta^2 = .03$).

For PPG, the same overall pattern emerged for HRV, but not for HR. Mean heart rate was significantly modulated by difficulty ($F(1, 25) = 10.01$, $p < .01$, $\eta_p^2 = .29$), with post-hoc analyses indicating higher values in the hard condition (session 1: $p = .0269$, $\eta^2 = .00$). HRV decreased significantly with difficulty ($F(1, 25) = 7.93$, $p < .01$, $\eta_p^2 = .24$), with post-hoc tests confirming lower values in the hard condition (session 1 : $p = .0175$, $\eta^2 = .01$).

For EDA, only phasic mean amplitude increased significantly with higher cognitive demand ($F(1, 25) = 14.11$, $p < .001$, $\eta_p^2 = .36$). Post-hoc comparisons confirmed greater phasic activity in the hard condition compared to the easy condition in both sessions (session 1: $p = .0434$, $\eta^2 = .13$; session 2: $p = .0049$, $\eta^2 = .14$).

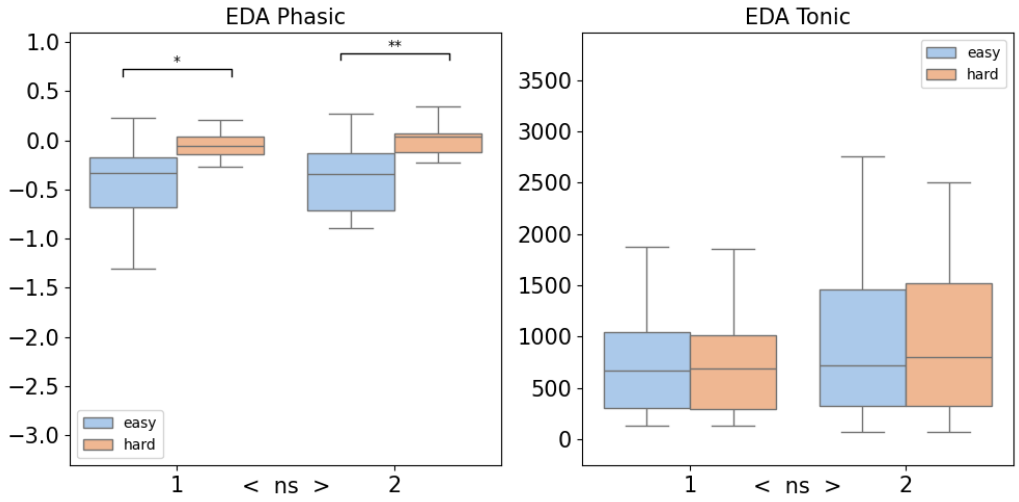


Figure 6.15: Results for the Sternberg task over EDA metrics. The session effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

### Cognitive Shifting

In the Cognitive Shifting paradigm, for RSME score, session had a significant effect ($F(1, 26) = 5.05$, $p < .05$, $\eta_p^2 = .16$), with lower values in session 2 than in session 1. The effect of difficulty was highly significant ($F(1, 26) = 84.21$, $p < .001$, $\eta_p^2 = .76$). Posthoc results confirmed higher effort ratings in the hard compared to the easy condition across both sessions (session 1: $p < .001$, $\eta^2 = .26$; session 2: $p < .001$, $\eta^2 = .29$).

For % correct, a significant effect of session was observed ($F(1, 26) = 9.72$, $p < .01$, $\eta_p^2 = .27$) with more correct answer in session 2 than in session 1. The effect of difficulty ($F(1, 26) = 31.77$, $p < .001$, $\eta_p^2 = .55$) and the interaction between session and difficulty ($F(1, 26) = 4.63$, $p < .05$, $\eta_p^2 = .15$) were also significant. Posthoc analyses confirmed that accuracy decreased significantly from easy to hard conditions in both sessions (session 1: $p = 0.0001$, $\eta^2 = .11$; session 2: $p = 0.0001$, $\eta^2 = .11$).

Similarly, for the F1 score, the effect of session was significant ($F(1, 26) = 10.25$, $p < .01$, $\eta_p^2 = .28$) with more correct answer in session 2 than in session 1. Difficulty also reached significance ($F(1, 26) = 24.59$, $p < .001$, $\eta_p^2 = .49$). Posthoc tests showed significant decreases in F1 score between easy and hard conditions in both sessions (session 1: $p = 0.0002$, $\eta^2 = .10$; session 2: $p = 0.0006$, $\eta^2 = .10$).

For RT, a strong session effect was found ($F(1, 26) = 17.29$, $p < .001$, $\eta_p^2 = .40$) with lower values in session 2 than in session 1. Difficulty effects were also significant ($F(1, 26) = 89.56$, $p < .001$, $\eta_p^2 = .78$). Posthoc analyses revealed slower responses in the hard condition compared to the easy condition in both sessions (session 1: $p < .001$, $\eta^2 = .48$; session 2: $p < .001$, $\eta^2 = .50$).

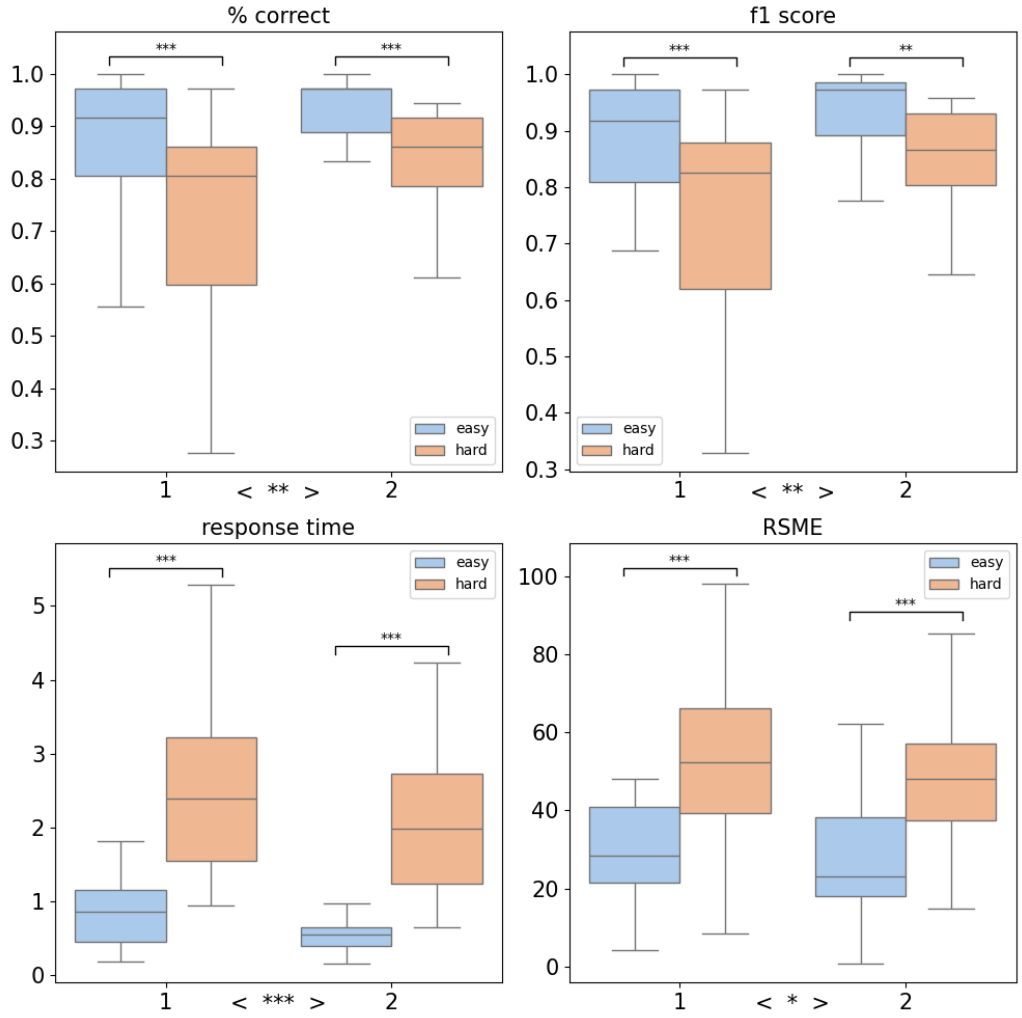


Figure 6.16: Results for the Cognitive Shifting task over the four non-physiological metrics. The session effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

For EEG, significant effects of difficulty were found for both central beta power and the central engagement index. For central beta, the effect of difficulty was significant ($F(1, 25) = 13.69$, $p < .01$, $\eta_p^2 = .35$), with post-hoc comparisons confirming reduced beta activity in the hard compared to the easy condition across both sessions (session 1: $p = .0101$, $\eta^2 = .01$; session 2: $p = .0082$, $\eta^2 = .01$). Similarly, for the central engagement index, difficulty

effects were significant ($F(1, 25) = 17.02$, $p < .001$, $\eta_p^2 = .41$), with post-hoc tests again showing smaller values in hard compared to easy conditions across both sessions (session 1: $p = .0058$, $\eta^2 = .01$; session 2: $p = .0053$, $\eta^2 = .01$).

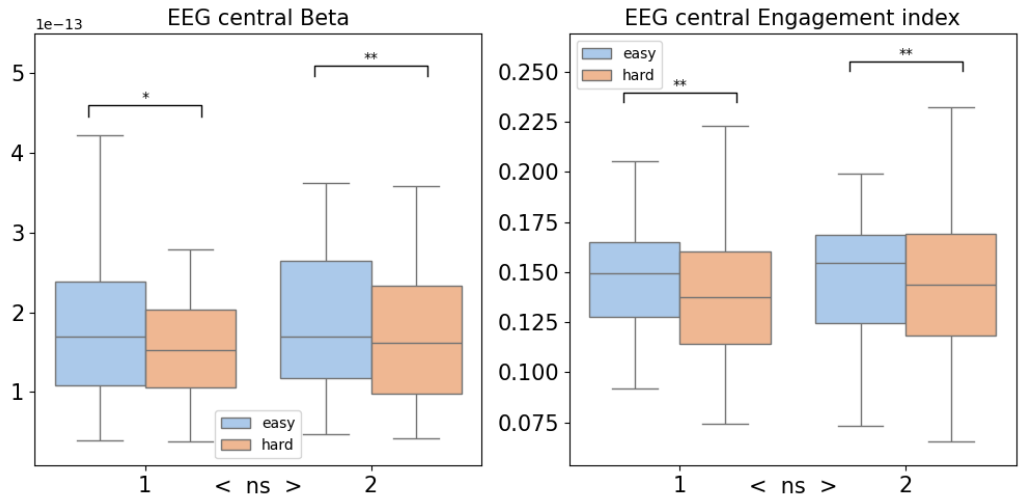


Figure 6.17: Results for the Cognitive Shifting task over EEG metrics. The session effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

For ECG, only HRV decreased with increasing difficulty ($F(1, 25) = 16.04$, $p < .001$, $\eta_p^2 = .39$). Post-hoc analyses confirmed lower values in the hard compared to the easy condition across both sessions (session 1: $p = .0018$, $\eta^2 = .07$; session 2: $p = .0349$, $\eta^2 = .02$).

For PPG, a similar but weaker effect was observed ($F(1, 25) = 6.78$, $p < .05$, $\eta_p^2 = .21$). Post-hoc tests indicated reduced variability in the hard condition (session 1 : $p = .0477$, $\eta^2 = .01$).

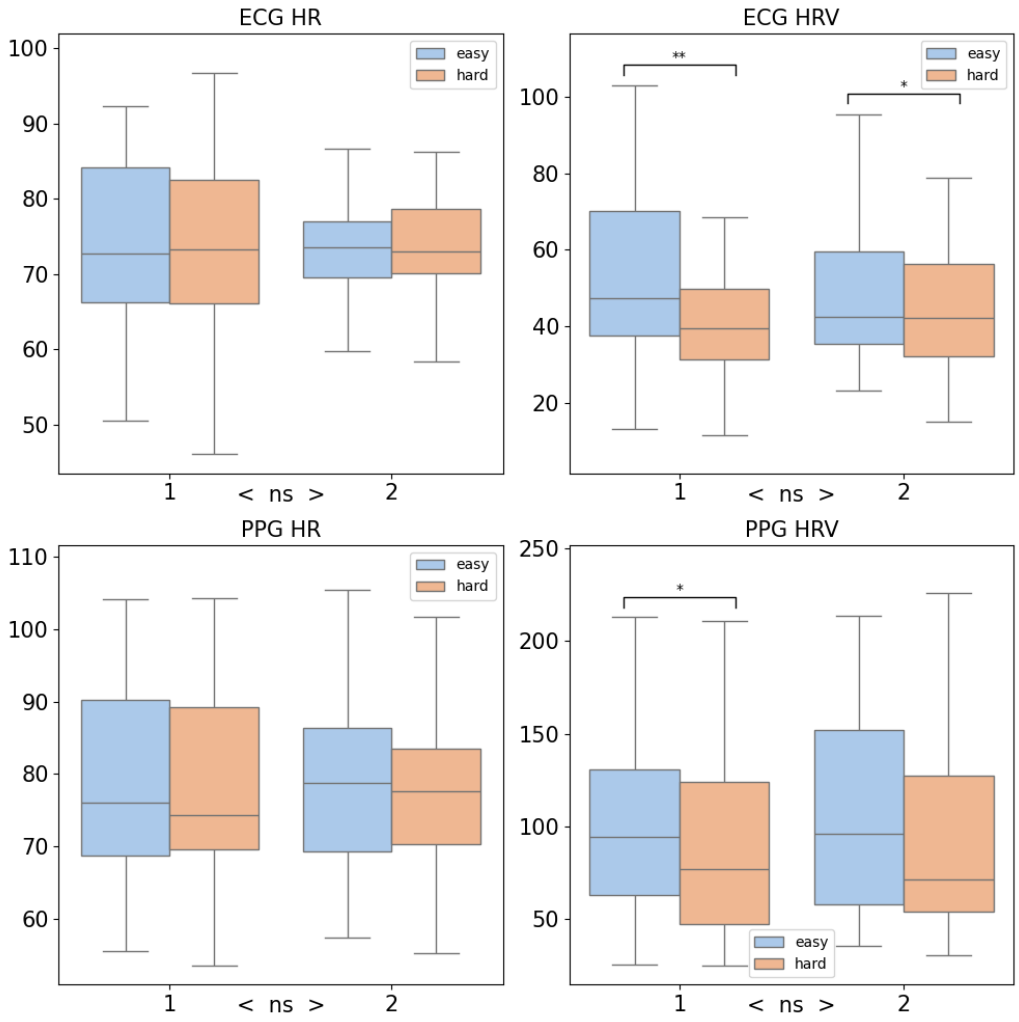


Figure 6.18: Results for the Cognitive Shifting task over ECG and PPG metrics. The session effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

For EDA, neither tonic nor phasic activity showed reliable modulation between easy and hard conditions, and no interaction with session was observed.

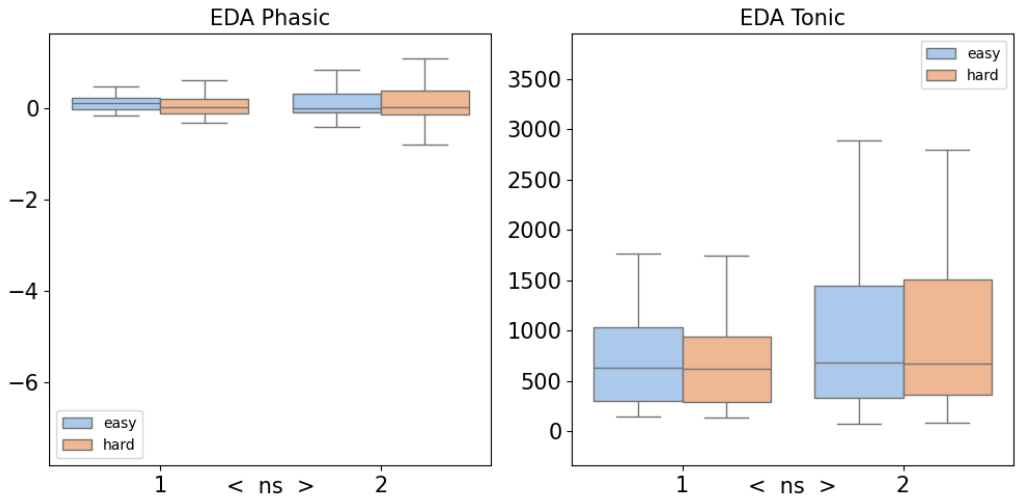


Figure 6.19: Results for the Cognitive Shifting task over EDA metrics. The session effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

### 6.4.5 Classification Results

A binary classification process was run for each paradigm. Classification performances reported for EEG are based on SVM classifiers with covariance matrices projected from the Riemannian space into the Euclidean space and then used as input features. While other classification pipelines have been tested (see Section 5.6.1), they did not reach similar performances during the 5-fold cross-validation and will only be presented during the general discussion of this manuscript. For the same reasons, only performances obtained with Random Forest classifiers are reported for ECG, PPG and EDA.

#### Sternberg

For EEG features in the Sternberg paradigm, the highest classification accuracy was achieved using beta-band activity, with a mean intra-session accuracy of 79.01% (79.25% in session 1, 78.77% in session 2). The "full" frequency range (1-100 Hz) followed closely with 77.95% on average (78.51% in session 1, 77.39% in session 2), while the "all" (1-30 Hz) range yielded 77.44% (77.90% in session 1, 76.97% in session 2). Lower performance was observed

for alpha (73.11%) and theta (69.91%) bands, though both remained above chance levels.

For ECG features, the best results were obtained with HRV features, reaching a mean accuracy of 56.79% (56.08% in session 1, 57.50% in session 2). Using all features produced comparable performance at 56.21% (55.21% in session 1, 57.21% in session 2), while HR alone yielded slightly lower but more variable accuracies across sessions (55.84% overall, 53.92% in session 1, 57.76% in session 2). All these results failed to exceed the adjusted chance level.

For PPG features, using all features achieved a mean accuracy of 54.86% (53.50% in session 1, 56.21% in session 2). Heart rate features alone yielded comparable results at 54.79% overall (52.57% in session 1, 57.00% in session 2), while HRV features slightly improved performance to 54.93% (53.06% in session 1, 56.80% in session 2).

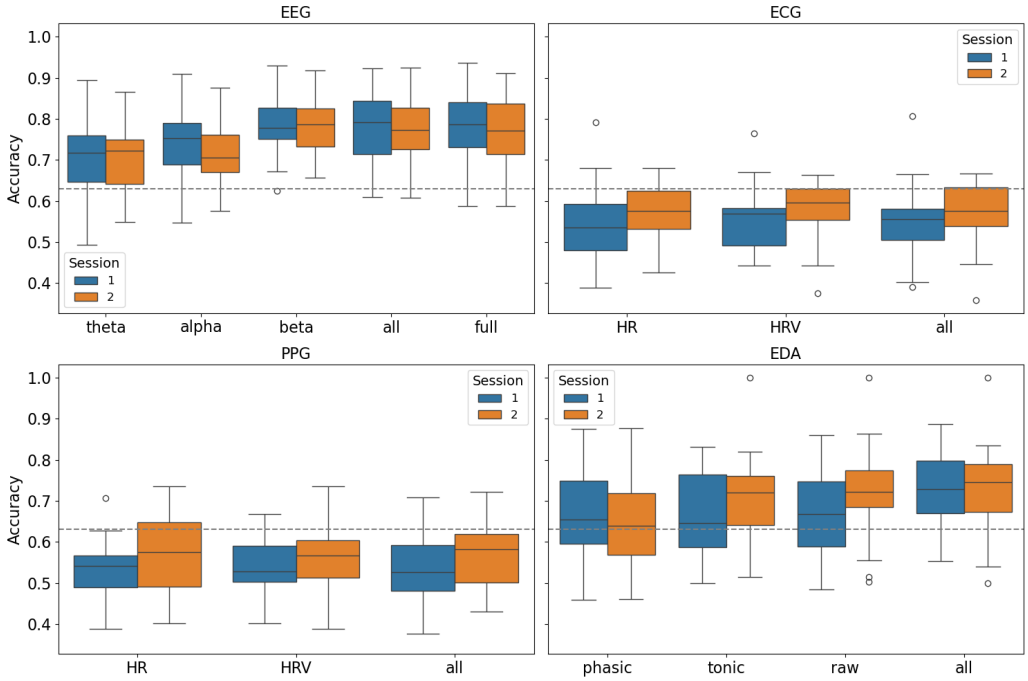


Figure 6.20: 2-class classification accuracy results for each feature of each signal. The dot-line represents the adjusted chance level for alpha = 0.05.

For EDA, using all features, the mean accuracy reached 73.03% (72.79% in session 1, 73.27% in session 2). Focusing on phasic activity alone yielded the lowest accuracy at 65.11% (66.64% in session 1, 63.57% in session 2). Raw

EDA signals achieved 69.48% mean accuracy (67.05% in session 1, 71.91% in session 2), whereas tonic features alone provided 68.51% overall accuracy (66.07% in session 1, 70.95% in session 2).

### Cognitive Shifting

For the Cognitive shifting paradigm, EEG features provided the highest classification performance. Using all frequency bands (1–30 Hz), mean accuracy reached 79.92% (79.76% in session 1, 80.09% in session 2). Restricting the analysis to beta frequencies improved performance to 81.95% (81.80% in session 1, 82.10% in session 2), while alpha frequencies yielded 75.60% (75.54% in session 1, 75.66% in session 2). Considering full frequencies up to 100 Hz achieved 80.94% mean accuracy (80.53% in session 1, 81.36% in session 2), and theta frequencies produced the lowest accuracy at 72.42% (72.71% in session 1, 72.14% in session 2).

For ECG features, overall performance was moderate. Using all features, mean accuracy reached 59.97% (60.82% in session 1, 59.13% in session 2), while HR features achieved 58.62% (58.99% in session 1, 58.25% in session 2). HRV features alone yielded 59.05% (60.43% in session 1, 57.67% in session 2).

PPG features showed slightly lower performance compared to ECG. Using all features, mean accuracy was 58.03% (58.42% in session 1, 57.64% in session 2). HR features produced 57.79% (58.07% in session 1, 57.52% in session 2), while HRV features alone yielded 57.03% (56.67% in session 1, 57.39% in session 2).

For EDA, combining all features resulted in 70.25% mean accuracy (70.10% in session 1, 70.40% in session 2). Phasic features alone provided the lowest accuracy at 55.78% (55.29% in session 1, 56.27% in session 2), raw EDA signals achieved 69.01% (68.01% in session 1, 70.00% in session 2), and tonic features reached 67.01% mean accuracy (65.26% in session 1, 68.75% in session 2).

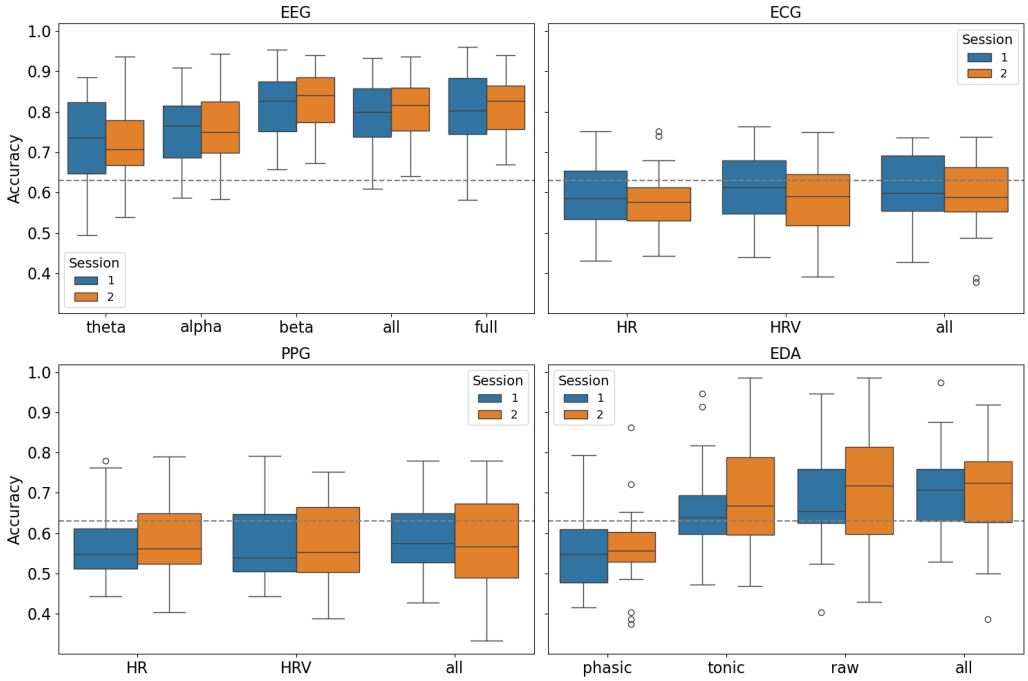


Figure 6.21: 2-class classification accuracy results for each feature of each signal. The dot-line represents the adjusted chance level for alpha = 0.05.

### 6.4.6 Main outcomes

In conclusion, this second campaign enriched the analysis by adding a physiological perspective alongside behavioral and subjective measures. While behavioral and subjective metrics consistently reflected workload, the interpretation of physiological signals was complicated by both session-related and task-specific variability. Among the different modalities, HRV emerged as the most stable marker across tasks, whereas EEG $\beta$ activity and engagement indices proved sensitive to workload but maintained consistency across sessions only in the Cognitive Shifting task. Conversely, EDA phasic responses tracked workload reliably in the Sternberg task across both sessions but failed to do so in Cognitive Shifting. Overall, these findings suggest that physiological markers of workload may be strongly task-dependent, underlining the need to identify robust indicators and to design models capable of generalizing across tasks and sessions in order to enhance the reliability of MSM in HRC.

## 6.5 RITHME: Comparison with a Physical Robot

### 6.5.1 Overview

Assessing whether a previously validated cognitive effort paradigm elicits comparable subjective, behavioral, and physiological responses when performed with a real robot requires replicating the task as faithfully as possible in both video-based and real-robot conditions. For this purpose, the Cognitive Shifting task from the database contribution was selected, as it was the most likely to be adapted as an actual HRC. Implementing it with the real robot, however, introduced minor adjustments: depending on the condition, participants had to interact either with the screen only (video clip) or with both the screen and a robot (real robot), as a computer was still required for answering. Moreover, instead of the laboratory-grade wet EEG used in the database study, participants wore a portable EEG system with dry electrodes, in order to better approximate real-world constraints and evaluate the feasibility of physiological monitoring in more ecological HRI scenarios.

### 6.5.2 Method

#### Task

The experiment relied on the Cognitive Shifting robotic task, implemented with the same structure and parameters as in the previous database contribution. It also comprised two levels of difficulty, manipulated through the cue: in the easier condition, the cue specified a single parameter to attend to, whereas in the more difficult condition, it indicated two parameters simultaneously.

For the video-based version, the task was strictly identical to the database contribution. Participants viewed sequences of robot actions presented as pre-recorded clips, with cues and response options displayed on the screen, and responses collected via keyboard input.

For the real-robot version, only the sequence of robot actions differed from the video-based implementation. The cue and target remained displayed on a screen, and responses were again given through a keyboard. The robot executed the same action scripts as those used to generate the videos in the earlier study, ensuring maximal comparability between the two versions. However, contrary to video clips, the physical robot was required to briefly maintain its rest position between signs, resulting in an additional inter-stimulus delay (less than 2 seconds).

This design ensured that the two task versions remained as similar as possible, with the only substantive difference being the stimulus modality.

### Measures

As in previous contributions, the Rating Scale Mental Effort (RSME) questionnaire was employed to capture subjective perception of cognitive effort [199]. Similarly, response time (RT) and two accuracy scores (% correct and F1 score) were used as behavioral measurements.

### Physiological Sensors

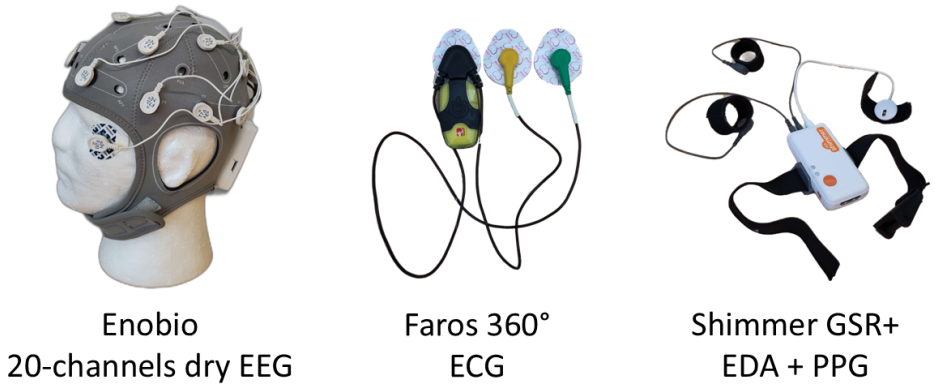


Figure 6.22: Illustration of the sensors used in the RITHME study for physiological data acquisition : an Enobio system with 20-channel, a Faros 360° device and a Shimmer3 GSR+ Unit.

In order to move closer to real-world constraints, wearability was here preferred over signal quality. Thus, brain activity was collected using a 20 channels Enobio (Neuroelectrics) with dry electrodes placed on a cap according to the 10/20 international system and sampled at 500 Hz. In addition, cardiac activity sampled at 1000 Hz was collected through electrocardiography using a Faros 360°. It was also collected through photoplethysmography, along with electrodermal activity, using a Shimmer3 GSR+ Unit sampled at 128 Hz.

### 6.5.3 User Study

#### Recruitment

20 participants were recruited via advertisements disseminated on the campuses of two universities. Their age, gender, education and professional situation are reported in table 6.4. In addition, participants' knowledge in robotics, physiology, and sign language was also collected on a 5-point likert scale.

| Gender | Highest Education | Professional Status | Knowledge |
|---|---|---|---|
| Female: 11 | Graduate : 12 | PhD Student: 9 | Robotics: 2.8 ± 1.1 |
| Male: 9 | Doctorate: 5 | Student: 5 | Physio: 3.2 ± 1.2 |
| | Undergraduate : 3 | Employed: 5 | Sign: 1.2 ± 0.6 |
| | | Unemployed: 1 | |

Table 6.4: Participant demographics and knowledge metrics for the 20 participants. Mean age was equal to 27.2 ± 4.0.

#### Protocol

Upon arrival, participants were welcomed and asked to complete an informed consent form and a demographic questionnaire. They then received a training phase consisting of three practice trials for each condition of each presentation modality. During this phase, participants were encouraged to ask questions, and the experimenter explicitly offered the possibility to repeat the training if necessary, though no participant chose to do so.

After training, the physiological sensors were set up: a dry EEG cap was placed on the scalp, and EDA/PPG sensors were attached to the fingers by the experimenter. For privacy and comfort reasons, the ECG electrodes on the torso were placed by the participants themselves, following a schematic guide. The main task then began and lasted approximately one hour. It consisted of 12 blocks of 12 trials each. At the end of each block, participants completed the Rating Scale Mental Effort (RSME) to provide a subjective measure of cognitive effort.

The block structure was organized so that participants always completed the two difficulty conditions of a given modality before moving to the next modality. The order of the two conditions within each modality was randomized. Consequently, modality switched every two blocks. To counterbalance presentation effects, half of the participants began with the video-based modality, while the other half began with the real-robot modality.

At the end of the session, physiological sensors were removed, participants received monetary compensation, and the experimenter conducted a debriefing, providing a more detailed explanation of the study's goals and addressing any questions raised by the participants.

This protocol has received approval from the institutional ethics review board of the University of Toulouse (project n°2025_1026).

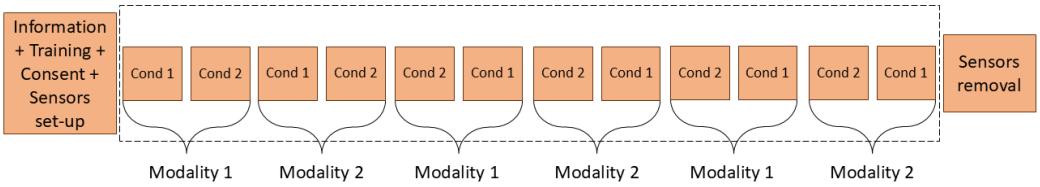


Figure 6.23: Experimental protocol for the RITHME study. Half of the participants began with the video-based modality, while the other half began with the real-robot modality.

### Analyses

Subjects were considered outliers if one of their behavioral results (accuracy, response time) diverged from others of more than 3 standard deviations in at least one condition. While this rule led to the rejection during results analysis of 0 subjects, one subject was still excluded due to sensor issues.

Data analysis has been performed using python, with the pingouin library for statistics [201]. Given the experimental designs, a two-way RMANOVA was applied to each metric for modality effect, difficulty level effect and interactions between them. In case of a significant effect, and after testing for normality, pairwise t-tests were then performed with Bonferroni correction.

For physiological signals, preprocessing and classification followed the same procedures as detailed in the previous chapter (see Section 5.4.3 and Section 5.6.2). The only adaptation concerned the choice of temporal windows, which was task-dependent. In the present case, the cognitive shifting task includes four videos per trial. Thus, ECG, PPG, and EDA signals were segmented into epochs spanning the four videos, yielding a temporal window of 14 seconds. Because of the faster dynamics of brain activity, EEG signals were instead epoched for each of these four videos separately, resulting in four times as many epochs as for the other signals but with a duration of 3.5 seconds each.

### 6.5.4 Statistical Results

For the RSME score, a significant effect of task difficulty was observed ($F(1, 18) = 67.78$, $p < .001$, $\eta_p^2 = .79$). Posthoc comparisons confirmed a higher RSME score in the hard condition compared to the easy condition for both stimulus modalities, with effect sizes of $\eta^2 = .40$ for the video and $\eta^2 = .34$ for the physical robot.

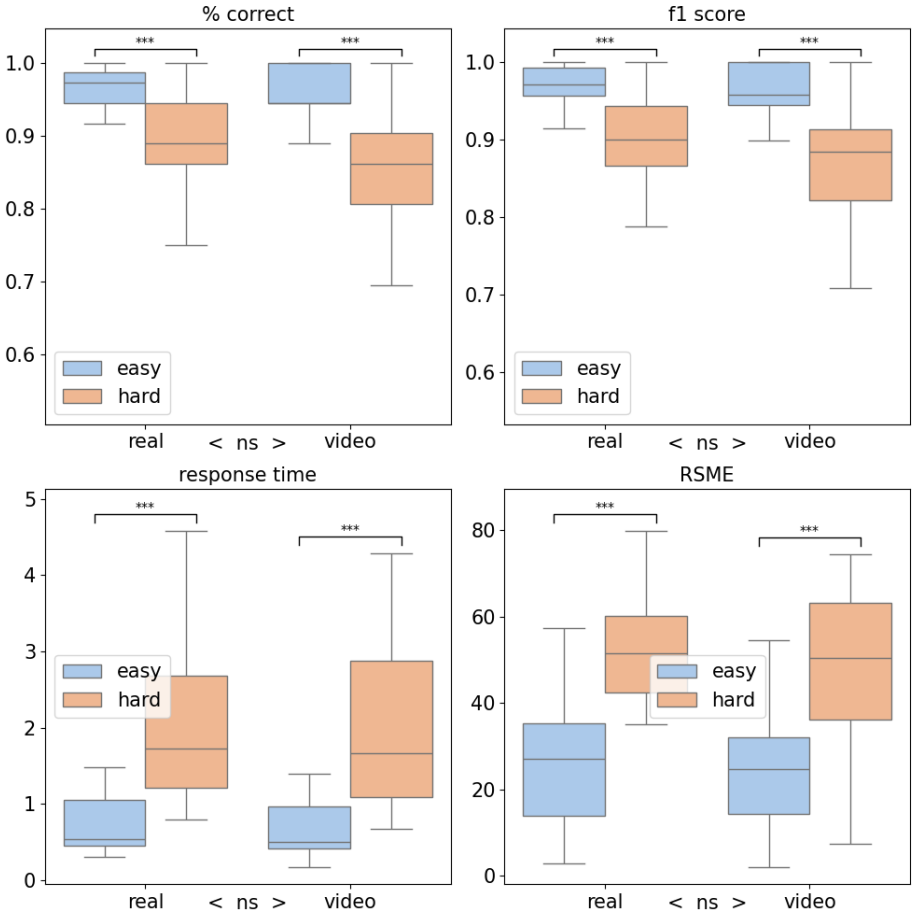


Figure 6.24: Results for the Cognitive shifting task over the four non-physiological metrics. The modality effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

For the %correct, the difficulty effect was significant ($F(1, 18) = 27.48$, $p < .001$, $\eta_p^2 = .60$). Participants performed more accurately in the easy

condition than in the hard condition across both modalities, with posthoc effect sizes of $\eta^2 = .25$ for the video and $\eta^2 = .34$ for the physical robot.

For the F1 Score, task difficulty was also significant ($F(1, 18) = 24.77$, $p < .001$, $\eta_p^2 = .58$). Higher F1 scores were observed in the easy condition compared to the hard condition for both video ($\eta^2 = .26$) and robot ($\eta^2 = .30$) modalities.

For RT, the timings were significantly longer in the hard condition compared to the easy condition ($F(1, 18) = 49.26$, $p < .001$, $\eta_p^2 = .73$). Posthoc tests confirmed slower responses in the hard condition for both video and physical robot modalities, with effect sizes of $\eta^2 = .40$ in both cases.

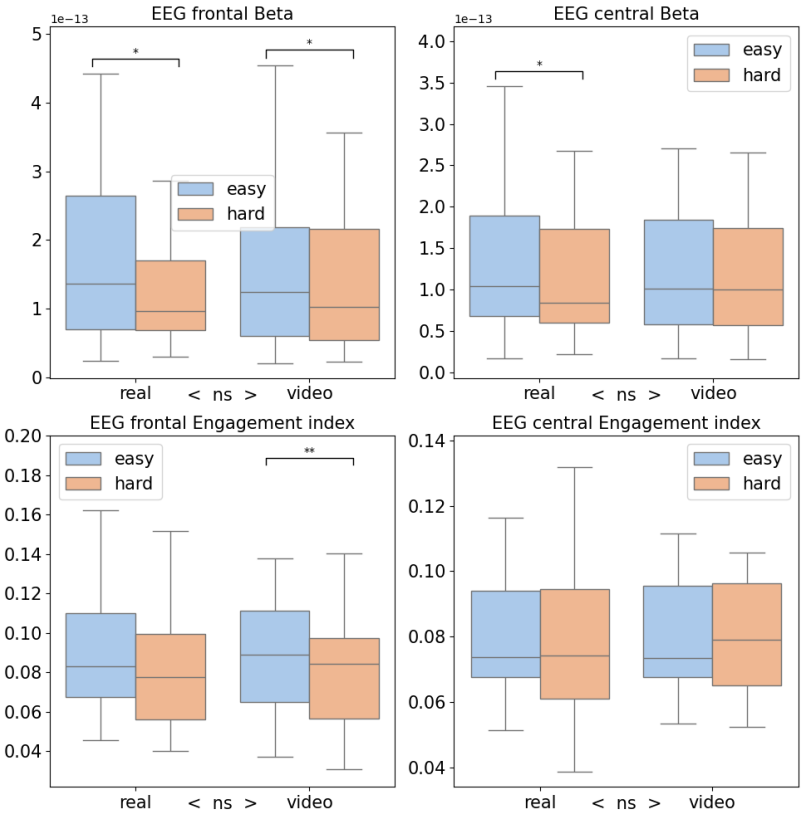


Figure 6.25: Results for the Cognitive shifting task over the EEG metrics. The modality effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

For EEG measures, task difficulty impacted both frontal and central re-

gions. Frontal beta power was significantly lower in the hard condition compared to the easy condition ($F(1, 18) = 10.19$, $p < .01$, $\eta_p^2 = .36$), with posthoc effect sizes of $\eta^2 = .02$ for video and $\eta^2 = .01$ for the physical robot. Central beta also exhibited a significant difficulty effect ($F(1, 18) = 5.4$, $p < .05$, $\eta_p^2 = .23$), although only with the physical robot, with lower values in the hard condition ($\eta^2 = .01$). Similarly, the frontal engagement index decreased with task difficulty ($F(1, 18) = 9.99$, $p < .01$, $\eta_p^2 = .36$; posthoc $\eta^2 = .01$), although only the video modality reach significativity .

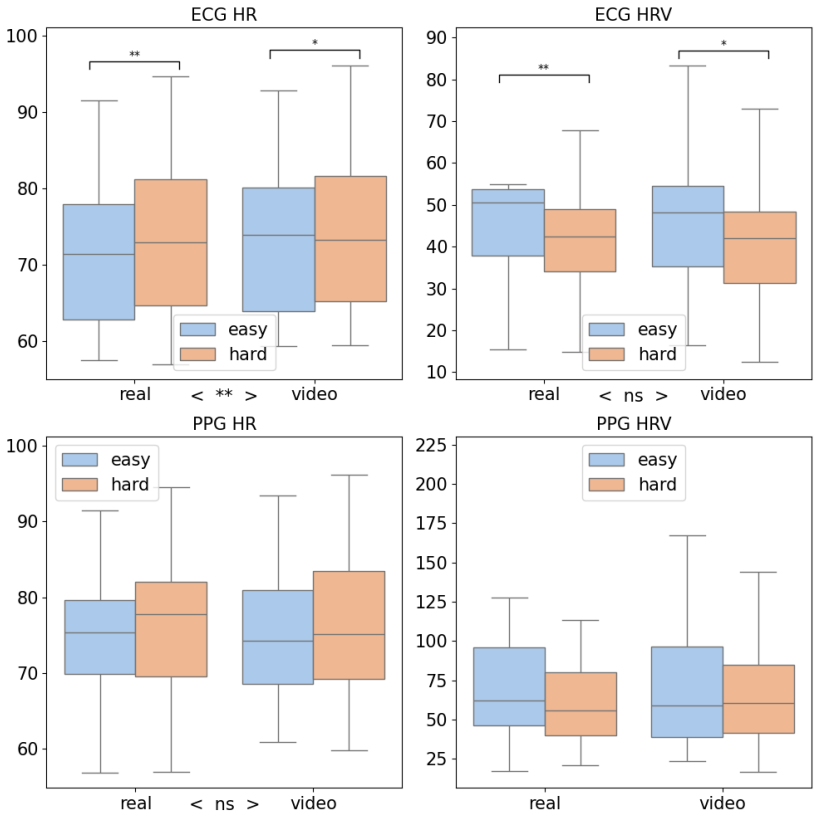


Figure 6.26: Results for the Cognitive shifting task over the cardiac metrics. The modality effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

For ECG, both modality and task difficulty significantly modulated heart activity. Mean HR showed a main effect of modality ($F(1, 18) = 9.02$, $p < .01$, $\eta_p^2 = .33$) and of difficulty ($F(1, 18) = 13.22$, $p < .01$, $\eta_p^2 = .42$). Post-

hoc comparisons confirmed higher HR in the hard condition compared to the easy one across both modalities (all $p < .02$, $\eta^2 = .01$). HRV also revealed a robust effect of difficulty ($F(1, 18) = 22.45$, $p < .001$, $\eta_p^2 = .55$), with post-hoc analyses showing a significant decrease from easy to hard conditions (all $p < .03$, $\eta^2 = .01$–$.03$).

For PPG, a significant main effect of difficulty was found for mean HR ($F(1, 18) = 5.17$, $p < .05$, $\eta_p^2 = .22$), indicating higher values in the hard condition. Although, post-hoc analyses did not reach significance for any modality.

For EDA, both phasic and tonic components revealed sensitivity to the experimental manipulations. Mean phasic activity showed a strong main effect of modality ($F(1, 18) = 15.79$, $p < .001$, $\eta_p^2 = .47$), as well as a significant interaction between modality and task difficulty ($F(1, 18) = 5.22$, $p < .05$, $\eta_p^2 = .22$). In addition, tonic activity exhibited a main effect of difficulty ($F(1, 18) = 5.59$, $p < .05$, $\eta_p^2 = .24$), with higher tonic levels in the easy compared to the hard condition, although no post-hoc contrasts reached significance.

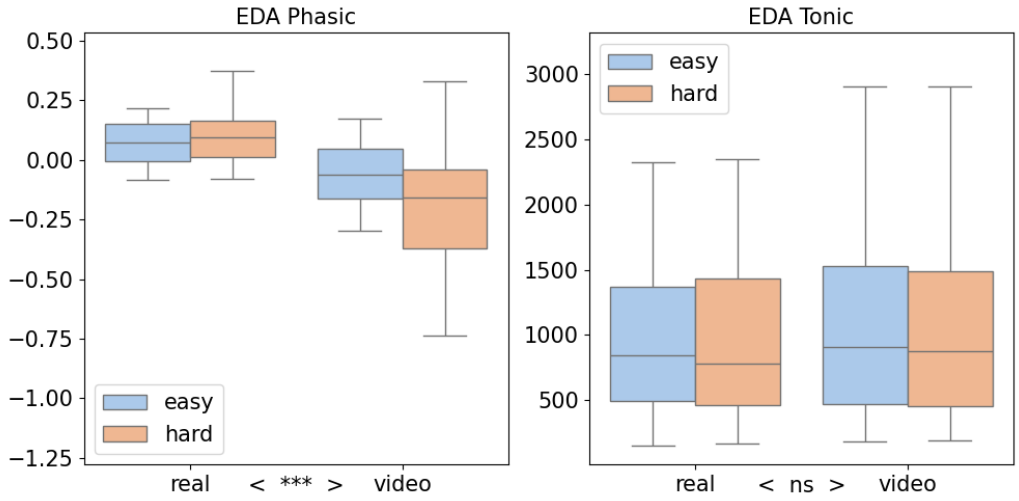


Figure 6.27: Results for the Cognitive shifting task over the EDA metrics. The modality effect is reported at the bottom of each figure while differences between groups are directly displayed at the boxplot level. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

### 6.5.5 Classification Results

Cognitive effort binary classification was run on the data. Classification performances reported for EEG are based on SVM classifiers with covariance matrices projected from the Riemannian space into the Euclidean space and then used as input features. While other classification pipelines have been tested, they did not reach similar performances and will only be presented during the general discussion of this manuscript. For the same reasons, only performances obtained with Random Forest classifiers are reported for ECG, PPG and EDA (see Figure 6.28).

For EEG, beta-band features achieved the highest classification accuracy with 78.79 % for the real robot and 80.65 % for the video modality. Using all frequency bands (1–30 Hz) resulted in 74.67 % for real and 77.00 % for video. Alpha-band features reached 72.30 % (real) and 73.96 % (video), full-band features (1–100 Hz) gave 72.35 % (real) and 74.36 % (video), and theta-band features achieved 70.56 % (real) and 70.65 % (video).

Classification performance based on ECG features was moderate. Using all features combined resulted in 64.84 % for the real robot and 62.98 % for the video. Only HR features achieved 62.96 % (real) and 61.74 % (video) while HRV features reached 65.53 % (real) and 60.09 % (video).

For PPG, all features combined reached 66.40 % for the real robot and 58.99 % for the video modality. Only HR features yielded 63.78 % (real) and 58.76 % (video) while HRV achieved 65.89 % (real) and 57.56 % (video).

Classification using EDA features reached above chance-level accuracy. Using all features combined achieved 72.95 % (real) and 74.82 % (video). Phasic features gave 58.77 % (real) and 57.22 % (video), raw signals achieved 69.22 % (real) and 72.58 % (video), and tonic features yielded 67.83 % (real) and 71.25 % (video).

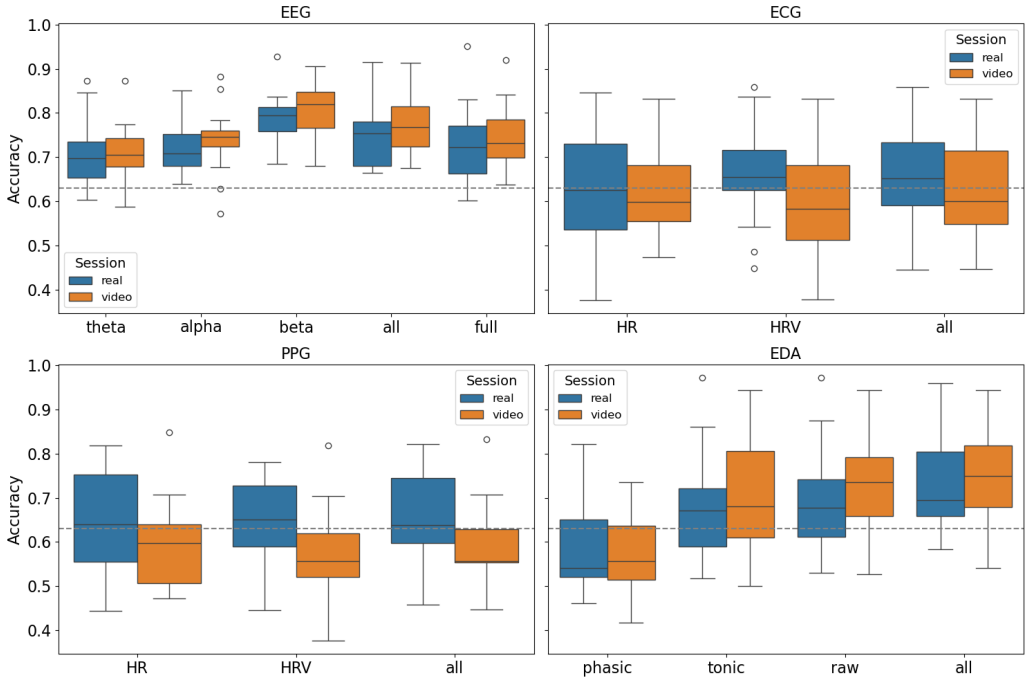


Figure 6.28: 2-class classification accuracy results for each feature of each signal (average across participants). The dot-line represents the adjusted chance level for alpha = 0.05.

### 6.5.6 Main outcomes

This third campaign demonstrated that robot-driven paradigms presented via video recordings can induce comparable behavioral and subjective levels of cognitive effort as direct interaction with a physical robot. Task difficulty consistently decreased accuracy, slowed responses, and increased perceived effort across both modalities, supporting the use of video stimuli as a valid, scalable alternative for cognitive induction. Physiological measures, however, revealed subtler differences: while EEG, particularly frontal beta activity, reliably tracked difficulty across sessions, cardiac and EDA signals showed modality-specific effects, with ECG and PPG classification performing better in real-robot conditions. These findings highlight both the potential of video-based paradigms for scalable cognitive effort experimentation and the need to investigate further the effect of physical presence on cognition.

## 6.6 Discussion

The overarching goal of this chapter was to investigate whether cognitive effort can be effectively induced by robot behavior itself, without relying on an additional secondary task. To this end, three protocols were conducted across complementary settings.

The first protocol aimed to establish that the robotic versions of the N-Back, Sternberg, and Cognitive Shifting tasks successfully induced cognitive effort. Across all three, increasing task difficulty consistently decreased accuracy and increased both subjective effort and response times, replicating classical findings from letter-based paradigms [186, 193]. Interestingly, in each task, robotic stimuli appeared to impose an additional layer of cognitive demand relative to letter stimuli. This effect is likely linked to the greater perceptual and cognitive load involved in interpreting dynamic robotic movements compared to static symbols. This extends earlier findings showing that stimulus modality [202] or quality [203] can substantially modulate cognitive workload. While this makes the robotic versions more challenging, it also highlights how using letter-based versions offer a limited insight on cognitive effort in real-world collaboration with robots. Nevertheless, small differences also emerged between the tasks. The N-Back results suggest that difficulty scales comparably across modalities, with only minor interactions. Yet, the repetitiveness of the task made it the less appreciated by participants. The Sternberg task showed a steeper increase in difficulty in its robotic implementation compared to the classical letter-based version, which typically employs a larger set of stimuli [193]. This divergence limits direct comparability with traditional paradigms but does not undermine the potential of this robotic variant. The Cognitive Shifting task, being novel, offered no benchmark for comparison, but produced clear and consistent differences across difficulty levels, particularly for robotic stimuli.

The second protocol extended the Sternberg and Cognitive Shifting tasks over two sessions and enriched the analysis with physiological data in addition to behavioral and subjective metrics. Consistent with previous findings, task difficulty reliably modulated cognitive effort, as reflected in decreased accuracy, slower responses, and higher subjective effort[204]. However, a session effect emerged across most behavioral and subjective measures, suggesting learning or habituation despite the initial training procedure. This highlights the need for careful consideration of repeated-measure designs, as adaptation across sessions may mask or dilute task-induced effects, although no interaction between session effect and difficulty effect were observed here. Physiological data provided further insights into the robustness of different modalities. For the Sternberg task, ECG HRV and EDA phasic component

consistently tracked difficulty [204, 140], whereas EEG central beta and PPG HRV did so only in the first session. In the Cognitive Shifting task, EEG central beta, EEG engagement index, and ECG HRV showed consistent sensitivity across both sessions, while PPG HRV again appeared session-dependent. Interestingly, the reliable modulation of $\beta$ activity by task difficulty stands in contrast with the literature, where $\theta$ and $\alpha$ oscillations are more frequently associated with cognitive effort [143]. This atypical finding may be explained by the strong motor processing required with these robotic stimuli.

When these features were used for classification, EEG signals clearly outperformed all others, achieving accuracies around 80%, with EDA second-best (around 70%) and cardiac measures hovering around chance level. This divergence between statistical effects and classification outcomes underscores an important methodological point: while ECG HRV can detect differences in difficulty, short windows (7s here) may be too narrow to capture its slower physiological dynamics [111]. In contrast, EEG and EDA, proved better suited for short-term classification.

The third contribution tested whether robot-driven paradigms induce similar levels of cognitive effort when presented via video recordings rather than through direct interaction with a physical robot. The subjective and behavioral results were consistent across modalities and with litterature [204]: increasing task difficulty reliably reduced accuracy, slowed responses, and increased perceived effort, with no significant differences between video and real-robot conditions. Thus, video stimuli could serve as valid and scalable alternatives to physical robots for cognitive induction, greatly facilitating experimental replication and accessibility. However, physiological measures painted a more nuanced picture. Both ECG HR and EDA phasic component showed significant modality effects, suggesting that while videos and real-robot similarly modulate cognitive effort, they are still impacting differently other aspects of the interaction, as stated in other studies [205]. Concerning difficulty effect, EEG once again emerged as the most reliable modality with frontal beta activity consistently reflecting difficulty across sessions, while central beta and the engagement index were sensitive in only one session. ECG HRV also showed significant sensitivity to task demands. Classification analyses reinforced these findings, with EEG achieving the highest performance (around 80 %), followed by EDA (around 70 % each) and, barely above chance level, by ECG and PPG (around 65 % each). Notably, ECG performed better here than in the database protocol, a difference plausibly explained by the longer temporal window used (14s instead of 7s), which better matches the slower dynamics of cardiac responses. In addition, cardiac measurements both performed systematically better with the real robot compared to the video modality, an insight that, coupled with the statisti-

cal effect observed on ECG HR and EDA phasic component, could require further consideration in a dedicated study.

Taken together, these three campaigns converge on several key conclusions:

1. First, robot-driven paradigms are capable of reliably inducing cognitive effort, replicating core findings from traditional tasks while offering increased ecological perspectives.

2. Second, while all signals showed some sensitivity to effort modulation, EEG and EDA emerged as the most reliable for classification, with cardiac measures showing promise but requiring longer windows.

3. Third, physiological variability remain a pervasive challenge, underlining the need for adaptive approaches such as transfer learning or calibration procedures for robust longitudinal monitoring.

Finally, the equivalence observed between real and video stimuli suggests a practical path forward for scaling experimental research, while also inviting future work to further investigate the meaning of the physiological changes observed when facing a physically present robot.

Ultimately, by focusing on robot-driven rather than dual-task paradigms, this body of work demonstrates the feasibility of embedding mental state induction directly within HRC scenarios. This not only avoids divided attention but also opens new possibilities for adaptive collaboration, where robots adjust their behavior in real time based on detected cognitive effort. Future studies should build on this foundation by refining task parameters, extending temporal analyses of physiological signals, and integrating these paradigms into interactive, dynamic collaborations where the robot actively modulates cognitive demand.

## 6.7 Availability of data and materials

The three tasks are available with the parameters set for user study and the results collected here:

**https://github.com/RIS-WITH/RITHM.git**

The code to generate robot behaviour is available here:
**https://gitlab.laas.fr/epiic/rithm_signs.git/**

# Chapter 7

# Contrasting with an other mental state

## 7.1 Motivation

Error-related potentials (ErrPs) have long been recognized as a robust neurophysiological marker of error perception [206, 207]. Their reliable detection and classification hold great promise for Human-Robot Collaboration (HRC), as they provide a direct window into the user's implicit evaluation of a robot's actions. Recent research has explored embedding ErrP-based feedback into reinforcement learning frameworks [70, 208, 209, 68, 149], enabling robots to refine their behavior dynamically based on whether the human collaborator perceives their actions as correct or erroneous. While these approaches have yielded promising performances in controlled laboratory environments, significant challenges remain before ErrP-based monitoring can be deployed in real-world collaborative contexts.

A first challenge lies in the temporal dynamics of error perception. In paradigms with discrete actions, ErrPs occur within a relatively well-defined time window following the action, which facilitates detection. In contrast, continuous robot actions create multiple opportunities for participants to detect an error, making it much harder to identify the temporal boundaries of the associated ErrP [210]. Thus, the first contribution of this chapter presents a complete pipeline for offline and online ErrP detection, originally developed for a classification competition and benchmarked on a public dataset. Beyond validating methodological choices, this contribution provides a baseline for the real-time feasibility of ErrP classification in HRC.

A second challenge stems from the heavy reliance on binary-choice paradigms, which oversimplify error perception in HRC. In naturalistic settings, error

processing is influenced not only by the presence of an error but also by its source [211, 212] and its saliency [213]. These factors highlight the need for a richer taxonomy of error types to better capture the complexity of collaborative work. To address this limitation, the second contribution proposes a novel HRI scenario in which errors emerge dynamically within a shared workspace, requiring participants to monitor and evaluate robot actions in real time. This paradigm goes beyond binary correctness and specifically examines participants' reactions to two distinct error types, combining EEG-based ErrP detection with subjective reports and behavioral measures to develop a more nuanced understanding of error perception in HRC.

## 7.2 Classifying Error

### 7.2.1 Overview

This approach uses EEG signals to detect the occurrence of errors deliberately introduced during flexion and extension with an orthosis device, as well as to estimate the timing of these errors. It has been developed and tested on the IntEr-HRI competition dataset [214].

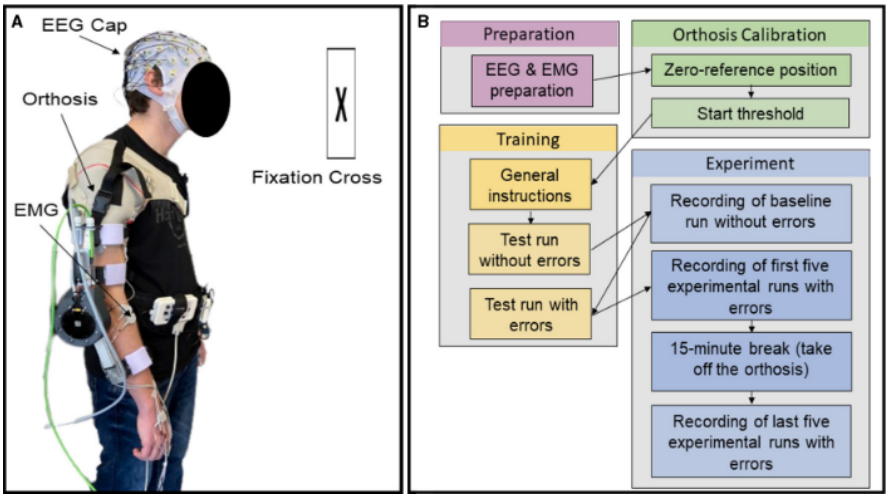


Figure 7.1: Setup (A) and protocol (B) of the data collection [214]

It used tactile feedback during flexion and extension movements of an active orthosis, where errors were defined as the device going briefly in the direction opposed to subject's movement (see Figure 7.1). During the main session, each participant performed 10 runs of 30 trials each, with six errors randomly introduced per run. To reduce artifacts in recordings, participants

were instructed to stand still and fixate on a cross during trials, which were initiated and terminated by verbal cues. A 15-minute break followed the fifth run, and participants were not informed of missed error reports during the experiment.

### 7.2.2 Data segmentation

In order to account for a realistic scenario, both training and test data have been loaded in a 900 ms buffer as a first step. A 15 Hz low-pass-filter and a common average reference (CAR) were then applied. Finally, a 500 ms sliding window with a leap of 100 ms allowed to segment data for classification.

A window was labelled as "error" if a physiological response to error (ERN) or to an odd event (P300) was expected in it [215]. As the P300b component has been noticed as particularly relevant to discriminate error condition according to training data first analysis, this response was selected as the main classification target. Thus, a window was labelled as "error" if the onset of a deliberately introduced error was present 400 ms before one point of the window.

It is noteworthy that, despite this focus on the P300b component, the classification pipeline still processes the whole 500 ms window. Thus, any other relevant physiological response present inside the window can be integrated by the model.

### 7.2.3 Pipeline

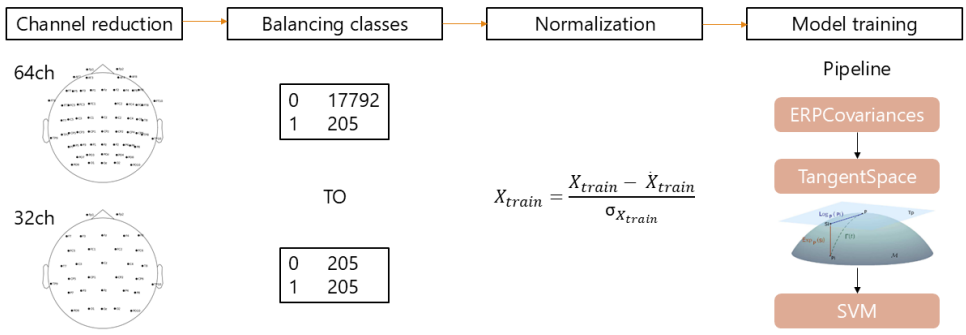


Figure 7.2: Illustration of the classification pipeline selected for this competition.

All the steps described hereafter were performed on 32 channels selected from the 10/20 system, as this reduced data shape and did not lower the performances compare to using the whole 64 channels (see Figure 7.2.) The whole processing pipeline is available on github (see 7.7)

Initially, the training set was prepared by randomly balancing the classes and normalized by dividing each sample by the standard deviation of the set. The testing set was subsequently normalized using the standard deviation of the training set. Classification was then carried out using a Riemannian approach, employing the pyriemann Python library [171]. Covariance matrices were computed and spatially filtered using a Riemannian variation of the common spatial pattern (CSP) algorithm [91]. These matrices were finally projected from the Riemannian space to the tangent Euclidean space in order to train an SVM model.

### 7.2.4 Sample selection

Due to the infrequent occurrence of errors, the tested model generated a substantial number of false positives, failing to accurately predict a sufficiently low number of errors. Consequently, an additional step was introduced to identify the most probable real errors among those predicted by the model (see Figure 7.3).

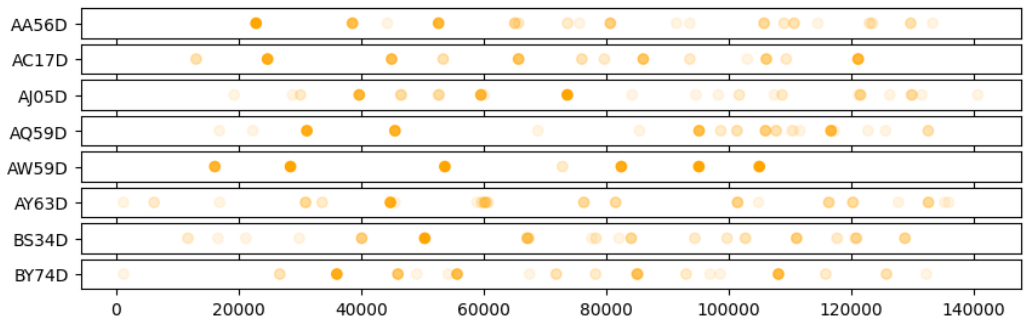


Figure 7.3: Results of the sample selection process for one run of each subject. High opacity reflects a large cluster of windows consecutively labeled as error.

The predicted errors were organized into clusters using a straightforward adjacency rule, where errors predicted on adjacent windows were grouped together. Clusters exceeding a predetermined threshold were then chosen. An iterative process was employed to define the most relevant threshold. As the expected number of errors was known during the competition, a subse-

quent processing enabled the retention of only the right number of evenly distributed clusters within the selected ones.

While the last part is a pure offline technique, it may be noteworthy that the used threshold could be learned from the training set.

Finally, the centroid window of each selected cluster was subjected to peak detection to identify the targeted physiological response. Although it was possible to calculate the typical latency between this response and the error onset to predict the error onset with a higher accuracy, we opted against this technique. The reason behind this decision was that it could potentially lead to predicting errors before their actual occurrence. Instead, we considered the sample of the detected peak as the representative error sample.

### 7.2.5 Pseudo-online detection

The approach was evaluated within each subject using a 10-fold cross-validation on the 8 provided datasets. In addition, because these data were not totally independent, special attention was paid to fold construction. Indeed, a classical leave-one-out 10-fold cross-validation could have included windows in the training set that overlap with windows from the testing set. Thus, for each fold, the 8 runs were randomly split into 6 runs for training and 2 for testing. This methodology was expected to introduce less bias and generalize well to the validation step on test data by following a similar approach.

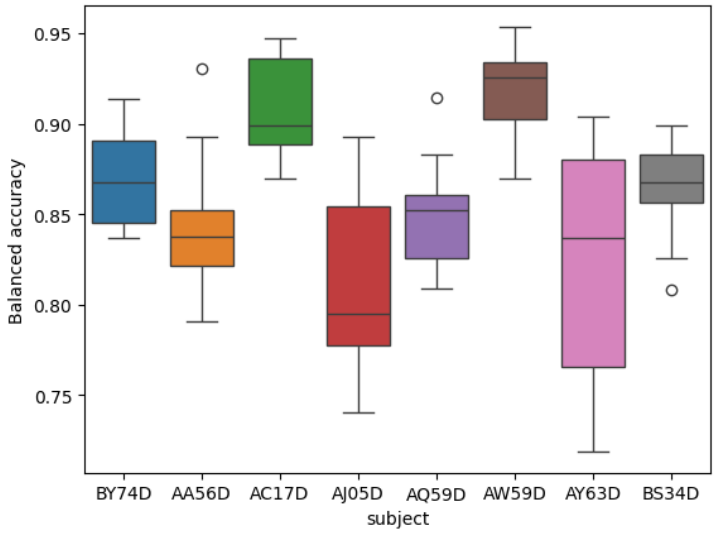


Figure 7.4: Offline classification performances obtained with the proposed pipeline, evaluated using a 10-fold cross-validation procedure.

Balanced accuracy scores ranged from 0.81 (subject AJ05D) to 0.92 (subject AW59D), with an overall mean of 0.86 ($\pm$ 0.04). These results provided a robust benchmark which allowed to reach third place during the competition.

## 7.3 DirectError

### 7.3.1 Overview

The objective of this study was to extend an established human–robot collaboration paradigm in order to examine how physiological monitoring can enrich our understanding of human responses to errors during interaction. The Director Task was selected to this end, a paradigm originally developed in psychology to study perspective-taking, and later adapted to HRI as a framework for designing robots with this capability.

The task was adapted to the experimental context, with a central modification being the controlled introduction of scripted errors. This design choice ensured that error occurrences could be systematically manipulated, allowing for rigorous comparisons in subsequent analyses. In addition, a simplified baseline task was incorporated at the end of the session to provide a reference condition, thereby facilitating the interpretation of physiological signals in relation to error processing.

The resulting user study focused on contrasting two categories of errors: those originating from perspective-taking failures and those being clearly erroneous. This distinction provided an opportunity to investigate not only the robustness of error-related responses, but also the extent to which different error sources elicit distinct effects on physiological, behavioral and subjective metrics.

### 7.3.2 Director Tasks

#### Psychology

The Director Task [216] is a widely used paradigm to investigate perspective-taking, a core component of Theory of Mind. Perspective-taking refers to the ability to infer what others can and cannot see, and to adapt one's own behavior accordingly. In this task, a participant and a director are placed on each side of a vertical grid (see Figure 7.5). To succeed, the participant must move objects according to the instructions received from the director. However, some cells are occluded on the director's side, and thus each agent has a different perspective on the grid.

This paradigm illustrates how inferring someone else's perspective allows to transmit unambiguous instructions with a reduced amount of information. By constraining the communicative context, the task provides a controlled way to examine how individuals resolve ambiguity, filter irrelevant information, and adapt to the knowledge states of others. It has been especially valuable in studying populations for whom perspective-taking and communication are critical challenges, such as anxious adolescents [217] or individuals with autism spectrum disorder [218].

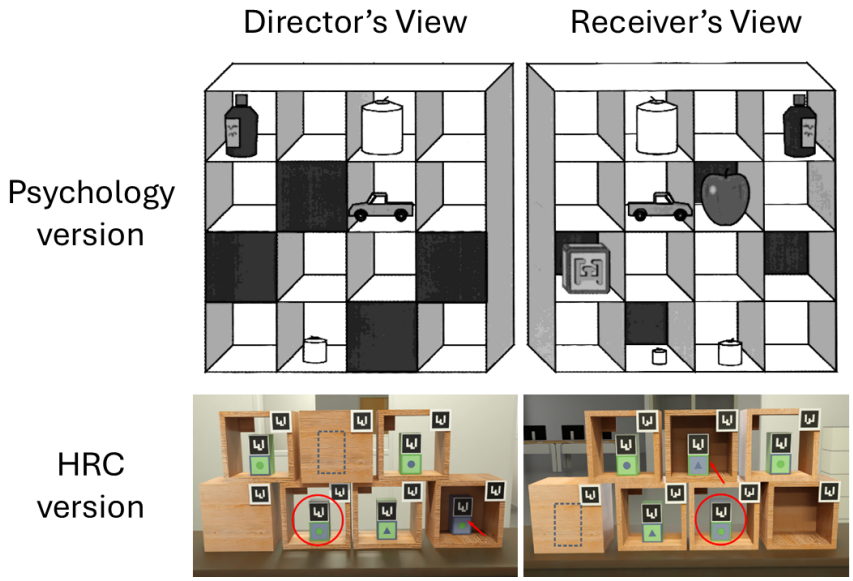


Figure 7.5: Illustration of the classic Director Task setup for both Psychology [216] and HRC [219].

### Robotics

In HRC, the Director Task has been adapted to preserve the central element of perspective-taking while introducing features that turn it into a richer collaborative scenario [219]. As in the psychological version, the task is performed on a vertical grid with occluded cells, and it involves two agents: a director who gives instructions and a participant who executes them (see Figure 7.5). The shared goal, however, is slightly different: rather than moving objects from one position to another, both agents must work together to remove a predefined set of cubes from the grid, in any order. This modification introduces temporal dynamics, as the configuration of the grid evolves with each removal, requiring both partners to constantly adapt their reasoning

and communication strategies.

A major difference with the psychological version is that both agents have occluded perspectives. In this HRI version, not only is the director unable to see all the cubes, but the participant also has a restricted view. As a result, both roles must actively reason about the other's perspective, creating a more balanced and reciprocal form of perspective-taking compared to the original paradigm, where only the participant has to infer the director's limited view. Moreover, to prevent reliance on simple geometric descriptions (e.g., "the leftmost cube"), instructions are restricted to the few visual features made available by the experimental setup and pointing is forbidden, ensuring that success depends primarily on perspective-taking rather than on spatial heuristics.

Another distinctive element of the HRI version concerns the stimuli themselves. Instead of generic objects, the grid contains cubes designed with four independent visual features: border color, cube color, the shape displayed and the shape's color. This multidimensional configuration provides a rich but controlled set of descriptive possibilities. Resolution requires careful reasoning about the other's restricted perspective to transmit a unambiguous yet minimal set of features. They allow for varied communicative strategies while limiting the vocabulary available and thus increasing the comparability between participants.

### 7.3.3 Methods

#### Controlling error

Such control requires to carefully balance task parameters as, to keep their impact, errors should be rare (between 20% and 30% of trials) and unpredictable (participants should not find a pattern in their occurrence). Yet, they should occur frequently enough to allow statistical analyses and appear homogeneously enough throughout the session to avoid any time-related bias.

#### Main Task

The DirectError task builds on the first HRI adaptation of the Director Task while introducing greater experimental control. In this version, only the participant acted as the director, while the robot executed the instructions. To stabilize the interaction, temporal evolution was removed: instead of removing cubes from the grid, the robot simply pointed at them. Thus, the participant's goal was to make the robot successively point to each cube visible to both agents within the grid (see Figure 7.6).



Figure 7.6: The grid is made of 4 x 4 cells excluding the 2 x 2 central ones. This spatial layout provide enough positions to point while keeping the grid compact and leaving the robot visible at the center. Among the 12 available cells, 9 are visible by both agents, 2 only by the robot and 1 only by the participant.

To do so, participants verbally described a target cube using any combination of five available features (see Figure 7.7). However, describing each cube with all five features would make the task trivial. To maintain the challenge and relevance of perspective-taking, participants were instructed to minimize their score, and thus use the smallest number of feature for each cube.

A typical trial began when the participant communicated an instruction containing an arbitrary number of features (see Figure 7.8). If the description matched exactly one cube, the robot pointed at it and verbally confirmed with a phrase such as "This is the cube you asked for". If the description matched no cube or too many, the robot asked to give a new instruction. To maintain a high number of successful trials despite the difficulty of the problem, cubes in the grid were carefully selected so that each could be uniquely identified with only two or three features. In addition, a tolerance

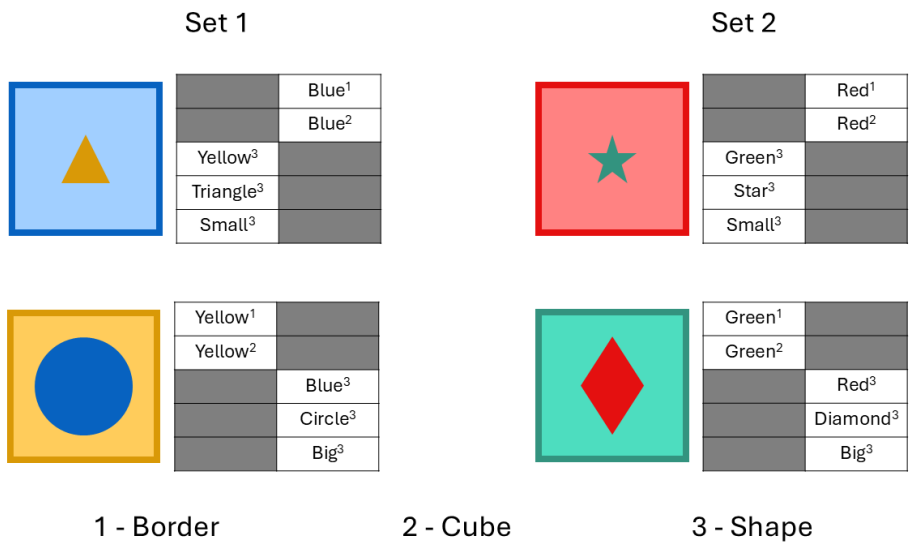


Figure 7.7: Four examples of possible cubes, the numbers on the corner explicit which part of the cube this feature belong to. Thus, the cubes display 5 features : border color, cube color, shape type, shape color and shape size. The latter was added compare to previous version to expand the number of possible combinations, increasing variability and the potential for communicative ambiguity.

system was implemented: if the description matched two or three cubes, the robot asked the participant to refine the instruction rather than start over, keeping the task fluent and limiting frustration.

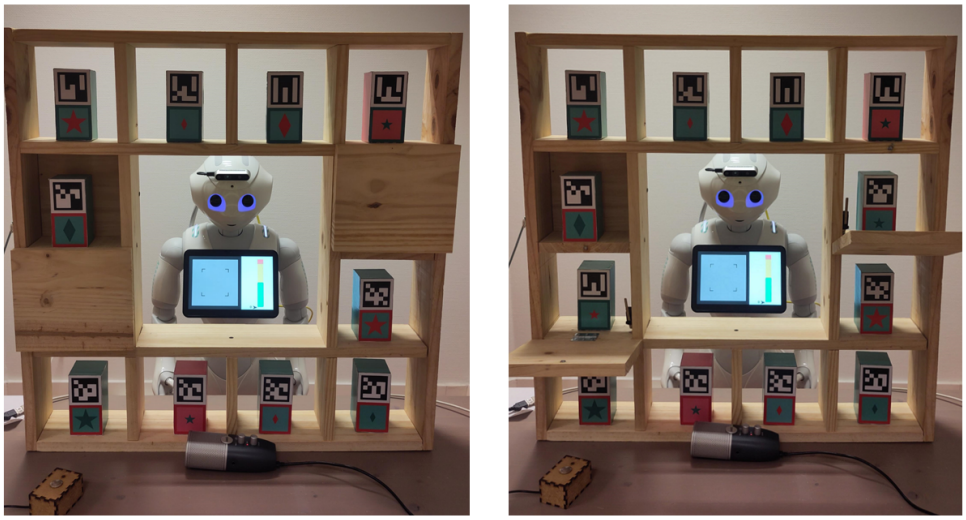

Figure 7.8: A typical setup from the participant's perspective. The left picture shows the classical view with occluded cells while the right view shows when they are exceptionally opened.

The central purpose of the task was to generate errors in a controlled yet naturalistic collaborative setting. Accordingly, three trial types were introduced:

- **Correct trials** where the robot pointed to the cube that matched the instruction and was visible to both agents.
- **Perspective failure trials** where the robot pointed to a cube that fit the description but was located in a cell visible only to itself, mimicking a failure to consider the participant's perspective.
- **Misinterpretation trials** where the robot pointed to a cube visible to both agents but inconsistent with the instruction, simulating a misinterpretation of the instruction unrelated to perspective taking.

To collect behavioral responses to these errors, two mechanisms were implemented. First, participants held a response box with a button and were instructed to press it as quickly as possible whenever they judged the robot's pointing to be erroneous. Second, they were given the option to open the cells visible only to the robot if the robot pointed toward them, with the timing of this action automatically recorded (see Figure 7.8). Crucially,

participants were informed that the definition of an error was entirely up to their own judgment: they were expected to press the button only if they genuinely perceived the robot's pointing as incorrect, and they remained free to decide whether or not to inspect hidden cells when the robot targeted them.

Finally, to support transparency and engagement, the robot displayed two forms of feedback during each block (see Figure 7.9). The first was a visual representation of the current instruction, shown as an incomplete cube that revealed only the features communicated so far. This aid helped participants monitor their descriptions in real time and reinforced that any errors should be attributed to the robot's decision-making process rather than misinterpretation of speech. The second was a score bar indicating the cumulative number of features used since the beginning of the block. By providing a clear performance indicator, this element triggered a sense of competitiveness and encouraged participants to formulate concise and efficient instructions.

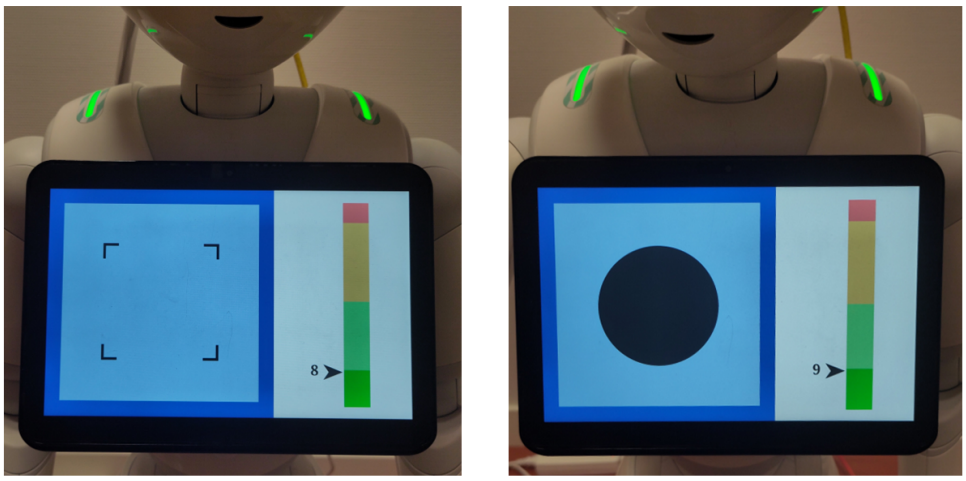


Figure 7.9: Example of feedback during the task. Left: the robot has already registered two parameters: cube color and border color. Right: the participant inputs an additional piece of information: the shape type. As a result, the interface now displays all three recorded parameters, and the participant's score increases by one.

Together, these design choices yielded a task that is highly controlled yet ecologically rich, enabling systematic investigation of how perspective-taking

failures, and more generally error, can be monitored during human-robot collaboration.

### Baseline Task

The baseline task (see Figure 7.10) was designed as a simple virtual grid-based game, adapted from prior research on ErrPs [220]. Participants controlled a green cube on the grid using the keyboard arrow keys, with the objective of reaching a red target cube within a maximum of ten moves. To systematically introduce errors, each keypress carried a 20% probability of producing an opposite movement, resulting in the cube moving away from the target instead of toward it. A trial was considered complete either when the green cube successfully reached the target or when the ten-move limit was reached.

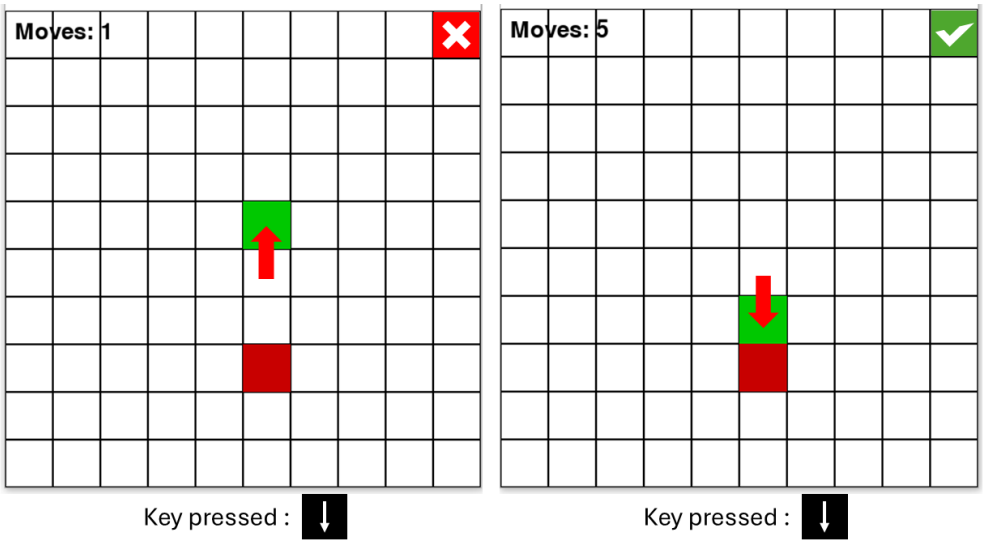


Figure 7.10: Example of the two kinds of trials during the baseline task. The left picture shows an erroneous trial while the right picture shows a correct trial.

As in the main experimental task, participants were instructed to press a response button whenever they detected an error. To ensure careful monitoring and prevent rapid, automated sequences of inputs, a fixed delay of 1 second was imposed between the keypress and the corresponding cube movement. This pacing encouraged participants to process each move individually,

thereby reinforcing error salience and facilitating clear neurophysiological responses.

#### Measures

Two behavioral measures were collected in both the main task and the baseline task: accuracy and response time (RT).

Accuracy was defined as the true positive rate (TPR), corresponding to the proportion of error trials in which participants correctly pressed the designated error button. It was computed as:

$$Accuracy(TPR) = \frac{TP}{TP + FN} \tag{7.1}$$

where $TP$ represents the number of correctly detected errors and $FN$ the number of undetected errors.

For the main task, TPR applied only to button pressed before inspecting the target cell. Consequently, in trials where the robot pointed toward a hidden cell, only button presses executed without first inspecting the hidden cube contributed to this metric.

Response time (RT) was defined as the temporal delay between the onset of the erroneous stimulus—either the initiation of the robot's pointing movement in the main task or the cube's displacement in the baseline task—and the participant's button press. This measure captures the speed with which participants detected and reported an error.

Subjective measures were also collected, not as primary outcomes, but to control for potential confounds such as fluctuations in participant reactivity or variations in attentional focus across time.

#### Physiological Sensor

As this study focuses on ErrPs, only brain activity was collected. The sensor was a 20 channels Enobio (Neuroelectrics) with dry electrodes placed on a cap according to the 10/20 international system and sampled at 500 Hz.

### 7.3.4 User Study

#### Recruitment

14 participants were recruited via advertisements disseminated on the campuses of two universities. Their age, gender, education and professional situation are reported in table 7.1. In addition, due to the nature of the task, colorblindness was an exclusion criteria.

| Gender | Highest Education Level | Professional Status | Knowledge |
|---|---|---|---|
| Female: 6<br>Male: 8 | Graduate: 8<br>Doctorate : 6 | Employed: 10<br>Student: 2<br>PhD Student: 2 | Robotics: 2.7 ± 1.3<br>Physio: 2.4 ± 1.5 |

Table 7.1: Participant demographics and knowledge metrics for the 14 participants. Mean age was equal to 35.4 ± 13.2.

### Hypotheses

This user study aims to investigate differences in responses to distinct error types at both the behavioral and physiological levels. Specifically, perspective-taking failures are expected to be more ambiguous and therefore elicit a more moderate response compared to misinterpretation errors. Based on this, the following hypotheses are proposed:

- Perspective-taking failure trials will exhibit a lower true positive rate than those involving misinterpretation errors.
- Perspective-taking failure trials will be associated with longer response times compared to misinterpretation trials.
- ErrPs will be elicited in error trials but not in correct trials.
- The amplitude and latency of ErrPs will differ between perspective-taking failure trials and misinterpretation trials.

### Protocol

Each experimental session followed a structured protocol (See Figure 7.11). Upon arrival, participants completed the informed consent form, while demographic information had already been collected during registration to save time. The experiment was introduced as a study on optimal communication, and the participants were told that several decision-making algorithms were being tested in a communication context, where errors might occur due to algorithmic limitations. The instructions were delivered following a carefully scripted explanation to ensure consistency across participants.

Before training, the robot's ability to understand participant instructions was verified using four typical sentences, with microphone parameters adjusted if necessary. Training then began: the experimenter guided participants through the first two cubes to ensure task comprehension, after which

they completed the remainder of the training grid autonomously. Only one misinterpretation error was introduced during training to familiarize participants with the concept of errors without overwhelming them. Training ended once participants had no further questions.

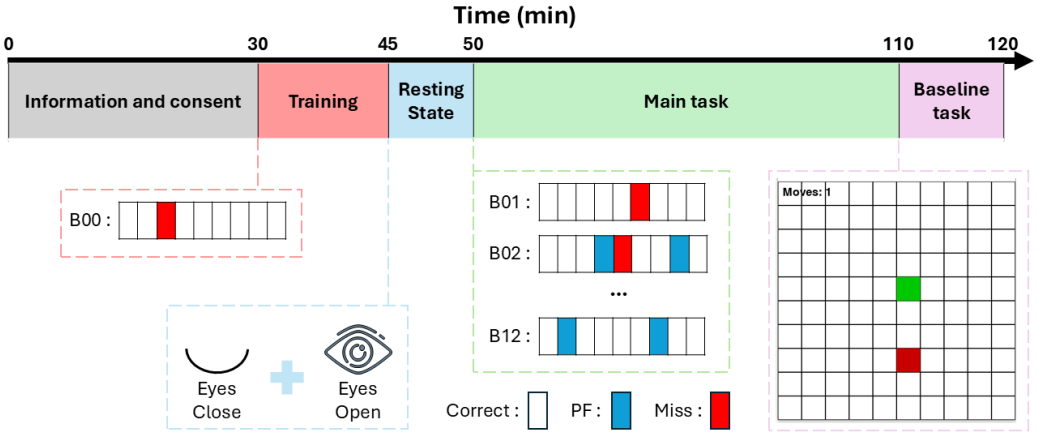


Figure 7.11: Timeline of the experimental protocol. Each partitioned row within the training and main tasks represents an experimental block: uncolored segments indicate correct trials, blue segments represent perspective failures (PF), and red segments correspond to misinterpretation errors (Miss). Since participants were free to complete the blocks in any order, the sequence of trials illustrated here is schematic and may not reflect the exact execution order.

Next, the EEG cap was set up, followed by a 5-minute resting-state recording consisting of alternating periods of 30 seconds with eyes open and 30 seconds with eyes closed, repeated five times. Participants then performed the main task, which comprised 12 blocks of 9 cubes each. In total, 108 cubes were described, including 24 error trials (12 misinterpretations and 12 perspective-taking failures, i.e., 22% of trials). Errors were distributed homogeneously across the task but not balanced per block, to avoid predictability. In case of an erroneous trial, participants were asked to describe the same cube again, and the robot then pointed to the correct cube. This repetition was justified to participants as the system switching to an alternative decision-making algorithm.

Between blocks, participants answered short subjective questionnaires on their own performance and their perception of the robot's performance. They were also offered the possibility to take breaks. To reduce habituation, the set of cubes alternated between blocks. After the main task, participants

completed a short post-questionnaire and then performed the baseline task for approximately 10 minutes. Finally, the EEG cap was removed, participants received monetary compensation, and they were debriefed about the true objectives of the study. The experimenter remained available to answer any remaining questions.

This protocol has received approval from the institutional ethics review board of the University of Toulouse (project n°2025_1095).

### Analyses

Two subjects were excluded due to technical issues, thus the analysis was performed on the remaining 12 subjects.

Data analysis has been performed using python, with the pingouin library for statistics [201]. Given the experimental designs, a t-test was applied to each metric.

For behavioral metrics, these comparisons were carried out between the two types of errors. For physiological metrics, contrasts were computed between error trials perceived as erroneous (TP) and correct trials perceived as correct (TN). Misperceived trials (FP, FN) where excluded to maximize interpretability at this stage by focusing on the clearest opposition. In addition, the analysis was again restricted to button presses performed during robot's pointing, thereby reducing potential noise from late or ambiguous responses.

Preprocessing and classification of the physiological signal followed the same procedures as detailed in the previous chapters (see Section 5.4.3 and Section 5.6.2). The only adaptation concerned the choice of temporal windows, which was task-dependent. Thus, in the main task, the analysis window encompassed the entire pointing action of the robot, starting from the initiation of the gesture and extending for 3 seconds, until the pointing movement was completed. This choice reflects the fact that an error could potentially be detected at any moment during this continuous interaction, requiring the full temporal span to be included. In contrast, in the baseline task, the analysis was restricted to the 1-second interval following the displacement of the green square. Here, error detection was assumed to occur almost instantaneously, making a shorter and more constrained window more appropriate.

## 7.4 Statistical Results

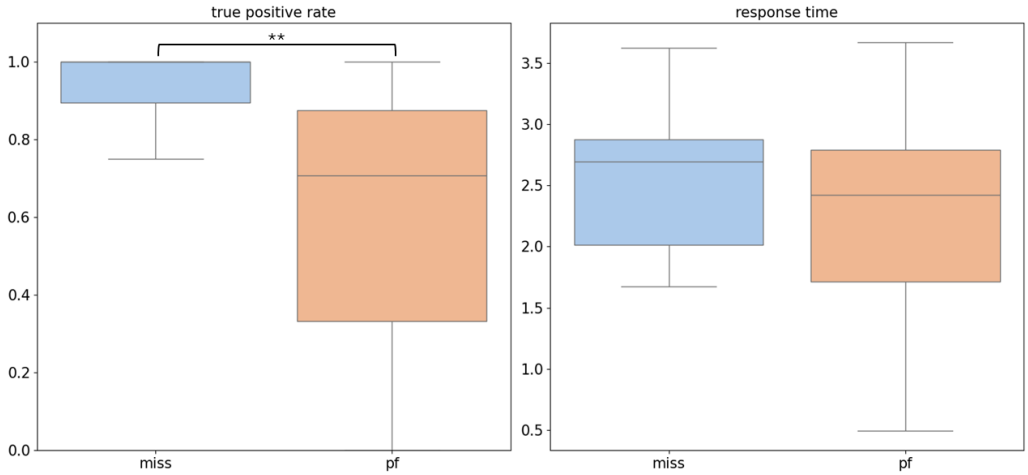


Figure 7.12: Statistical comparisons between the two error types, misinterpretation (miss) and perspective failure (pf), over the two behavioral metrics. Significativity is reported as star symbols with the following rule: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

Only the analysis of the true positive rate (TPR) revealed a significant effect of the condition ($T(11) = 3.17, p < .01, d = .24$). This indicates that participants were more likely to label a trial as containing an error when they were able to see the pointed cube, than when they were not.

## 7.5 Classification Results

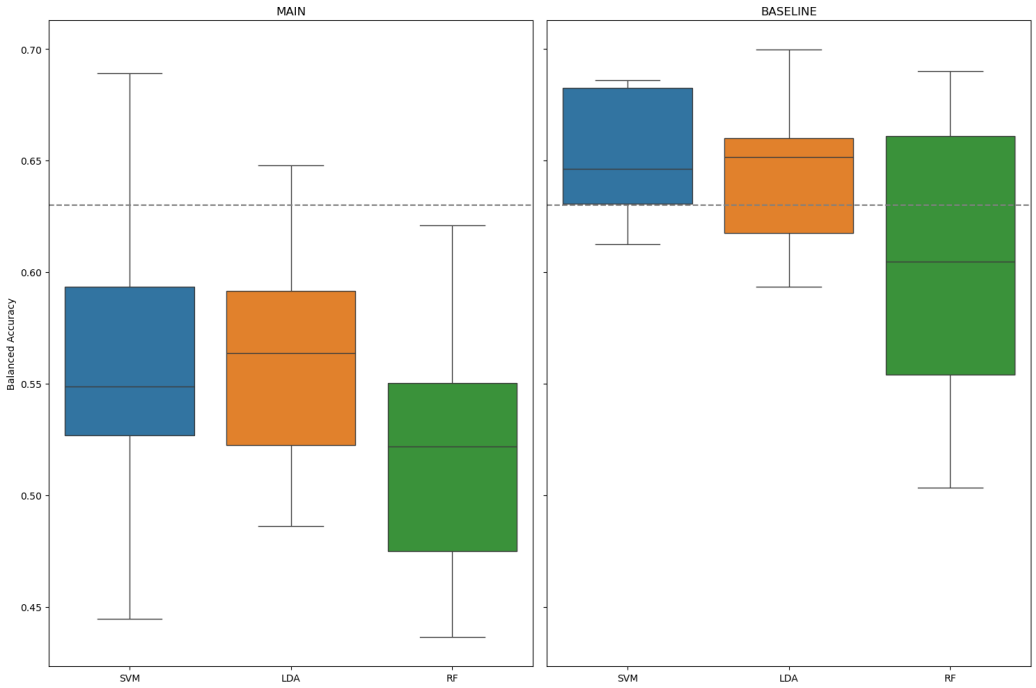


Figure 7.13: Classification results for the DirectError task over the two tasks. Chance level is reported as a dotted line.

For classification, only SVM classifiers in the baseline task achieved balanced accuracies above the corrected chance level. Specifically, SVM reached an average balanced accuracy of 65.2%, outperforming LDA (64.2%) and RF (60.1% which remained closer to chance level. In contrast, none of the classifiers achieved sufficient performance in the main task, where average balanced accuracies hovered around chance level (56.0% for LDA, 52.0% for RF, and 56.0% for SVM).

## 7.6 Discussion

The present chapter investigates error perception as a distinct mental state, shifting the perspective away from cognitive effort toward the neurophysiological signatures of error-related processes and their potential role in HRC. Whereas cognitive effort monitoring relies on the continuous assessment of oscillatory brain activity, error perception requires a focus on the temporal

dynamics of neural signals, capturing the event-related patterns characteristic of ErrPs. This contrast highlights the complementary methodological challenges associated with both approaches and broadens the scope of MSM in HRC.

The first set of results, obtained using a publicly available ErrP dataset, illustrated both the promise and limitations of error monitoring for online applications. Offline classification yielded strong performances, consistent with the robustness of ErrPs reported in prior literature [207]. However, pseudo-online analyses revealed important challenges, most notably a high false-positive rate, even with accurate offline classifiers. Applying temporal smoothing through clustering strategies mitigated these issues and was sufficient to reach third place in the competition with non-deep learning models. Still, recent advances in deep learning for ErrP detection [221, 222] highlight promising directions to further improve performance. These results also illustrate how open competitions and shared datasets can accelerate methodological progress at the community level.

Building on this foundation, the DirectError campaign collected a richer dataset during both a complex HRC scenario and a simpler baseline task. Preliminary behavioral results suggested that participants differentiated between the two error types. Specifically, perspective-taking errors were less frequently labeled as erroneous compared to misinterpretation errors, despite both worsening subject's score. This suggests that participants engaged in a more careful evaluation process for perspective-taking errors, often checking the hidden cell before committing to a judgment. These findings reinforce the idea that error perception in HRC goes beyond binary detection, needing a better characterization of context and saliency of the error [211]. Further work should examine the stability of this effect across time and individuals to better understand its variability and generalizability.

Neurophysiological classification, however, yielded more modest outcomes. In the main HRC task, classifiers did not perform above corrected chance levels, likely reflecting the complexity of capturing reliable ErrPs in continuous, naturalistic settings, where errors can be perceived at multiple moments [210] and signals are contaminated by motor activity [72]. In the baseline task, only SVM achieved marginally above-chance performance. A partial explanation may lie in the use of dry EEG systems, which, while offering practical advantages in wearability, can reduce signal quality [76]. Nonetheless, these results should be considered preliminary, and several avenues remain open for improvement, including the refinement of preprocessing pipelines, optimization of temporal windows, and exploration of more advanced classifiers.

If classification in the baseline task can be strengthened, it could serve as a training ground for cross-task transfer, enabling models trained on baseline

data to detect error responses in the main HRC task in a pseudo-online fashion. Achieving such transfer would mark a significant step forward, allowing error perception to be monitored not only through explicit markers such as button presses but also via implicit neural responses. This would enrich HRC with a more objective, multimodal account of human evaluation, requiring only a lightweight baseline calibration task alongside the main collaborative scenario.

Finally, while cognitive effort monitoring offers a continuous view of the user's sustained mental state, monitoring error perception demands finer temporal precision, which makes EEG the sensor of choice over other physiological modalities. This approach comes with additional requirements: high data quality is essential to reliably capture the specific temporal patterns of ErrPs and to account for discuss pattern variations across stimuli. At the same time, ErrP-based monitoring provides a unique advantage in terms of interpretability. Unlike frequency-domain measures of cognitive effort, which may be hard to grasp intuitively, the amplitude and timing of event-related potentials reflect direct neural responses. Taken together, these characteristics make error perception monitoring a challenging yet highly promising direction for HRC [149, 68, 67], marked by methodological constraints distinct from those of cognitive effort, but offering the potential for rich and transparent insights into human evaluation of robot actions.

## 7.7 Availability of data and materials

The code from the competition is available at https://github.com/mathiasrihet/IntEr-HRI.git.

# Part III

# General Discussion

# Chapter 8

# Discussion

This work investigated the use of physiological measurements for MSM in HRC. By focusing on cognitive effort and surprise as targeted mental states, it aimed not only to evaluate its feasibility but also to reflect on the methodological challenges that currently limit progress in the field. This general discussion synthesizes the contributions of this work, situating them within the broader context of HRC research. It emphasizes the central role of experimental design, compares solutions tested in this thesis, tries to identify stable biomarkers and classification pipelines through all the data collected, and identifies directions for future work.

## 8.1 A PhD about experimental design

The guiding principle of this PhD is that good experimental design is the starting point of reliable MSM in HRC. Without solid paradigms, physiological data cannot be meaningfully interpreted, no matter how sophisticated the subsequent processing or modeling. This observation is particularly pressing given the early stage of the field: while psychology and neuroscience have developed canonical tasks to study cognitive processes in controlled yet reproducible ways [184, 189, 216], HRC still lacks equivalent benchmarks [223, 224].

This absence creates a twofold problem. First, researchers attempting to integrate physiology into HRC often need to design their experiments from scratch, re-discovering basic pitfalls such as session effects, artifacts caused by robot motion, or the limits of wearable sensors. Second, the lack of common paradigms makes it nearly impossible to compare results across studies or build cumulative knowledge.

By foregrounding experimental design, this thesis thus positions itself less

as a contribution to algorithmic innovation and more as a contribution to the scientific infrastructure necessary for the field to mature. The experiments reported here explore not only what physiological signals can reveal about human states in HRC, but also how these signals can be elicited, recorded, and benchmarked.

| | **EPIIC** | **RITHMD** | **RITHME** | **DirectError** |
|---|---|---|---|---|
| **Mental State** | Cog Effort | Cog Effort | Cog Effort | Surprise |
| **Subject** | 22 | 26 | 19 | 12 |
| **Session** | 2 | 2 | 1 | 1 |
| **Trial (per Session)** | 60 | 144 | 144 | 132 |
| **Condition** | Easy/Medium /Hard | Easy/Hard x Paradigm | Easy/Hard x Modality | Correct/Error1 /Error2 |
| **Trial (per Condition)** | 20 | 36 | 36 | 108 (Correct) 12 (Error) |
| **EEG** | 19 Dry | 32 Wet | 19 Dry | 19 Dry |
| **ECG** | No | Yes | Yes | No |
| **PPG** | Yes | Yes | Yes | No |
| **EDA** | Yes | Yes | Yes | No |

Table 8.1: Summary table of the four experimental campaigns ran with physiological signals during this thesis. The number of subject reported is the number analyzed rather than the number recruited.

## 8.2 Comparing experimental designs

A central contribution of this PhD lies in the development of multiple experimental designs tailored to investigate MSM in HRC (see Table 8.1 for a summary). These designs showcase diverse ways in which physiological measurements can be integrated into HRC scenario, providing meaningfull insights into human cognitive processes. They are here summarized into three main categories, with a discussion of their respective advantages and limitations.

### 8.2.1 Dual-task paradigms

Dual-task paradigms introduce a secondary task alongside the main HRC scenario to modulate the mental state of interest, as in the EPIIC campaign (see Section 5.3). Their main advantage lies in their simplicity: they leverage well-established paradigms that are straightforward to design, implement, and replicate, making them particularly useful for rapid prototyping and initial exploration of new research questions [53, 180, 181]. However, this simplicity comes at a cost. Because the secondary task is artificially imposed, participants' behavior may deviates from what would be expected in a natural collaboration, limiting ecological validity [182, 183]. Furthermore, since the robot's actions play only a marginal role in shaping the induced mental state, these paradigms offer little potential for adaptive interaction. In short, dual-task designs are valuable stepping stones for early investigations into mental state monitoring in HRC, but their applicability to more ecological and adaptive settings may be limited.

### 8.2.2 Neuropsychology-inspired HRC scenario

Neuropsychology-inspired tasks adapt well-established paradigms, such as Sternberg [189] or N-back [184], into collaborative settings where the robot itself becomes the stimulus (see Sections 6.3, 6.4, 6.5). This approach achieves a balance between theoretical rigor and ecological relevance, grounding manipulations in robust cognitive frameworks while embedding them in the dynamics of HRC. A major advantage is the ability to systematically vary task condition while moving toward a naturalistic setting.

The trade-off lies in the effort and time required to carefully adapt and validate these paradigms. Such designs cannot leap directly into realistic HRC scenarios; they must progress step by step, beginning with controlled laboratory setups and gradually increasing in complexity. As a result, early implementations may appear dedicated primarily to mental state monitoring rather than to authentic collaboration, narrowing the scope of scenarios that can be investigated. Yet, this mirrors the trajectory of any MSM protocol, where simplified paradigms are essential stepping stones for uncovering fundamental cognitive mechanisms before extending them to richer real-world contexts [118, 225]. Similarly, neuropsychology-inspired HRC tasks enable researchers to formulate strong, theory-driven hypotheses grounded in well-validated paradigms, and then extend these designs step by step toward ecologically valid collaboration.

Although this incremental process is demanding, it may offer one of the most promising pathways for advancing MSM in HRC. By ensuring that

each step is both theoretically grounded and progressively more realistic, these tasks can serve as a bridge between cognitive neuroscience and applied robotics, helping the field grow in both methodological robustness and ecological relevance.

### 8.2.3 Using a baseline task

HRC scenarios with a baseline task combine the benefits of classical paradigms and naturalistic collaboration, without forcing participants to manage them simultaneously (see Section 7.3). In this design, the baseline task is carried out separately from the main HRC scenario, allowing researchers to gather clean, controlled data alongside complex, ecologically valid measures from the same participant within the same session. This setup enables direct within-subject comparisons between baseline and HRC conditions strengthening interpretation and reducing the impact of physiological variability across sessions.

However, an important limitation remains: cognitive, and thus physiological, responses often differ substantially between tasks [154]. This task-to-task variability can undermine the transferability of models and limit the immediate effectiveness of the approach [151]. In other words, while this design provides the tools necessary for cross-task analyses between a well-established reference paradigm and a realistic HRC scenario, it does not guarantee that these analyses will overcome the variability gap. Compared to neuropsychology-inspired designs, paradigms using a baseline-task allow researchers to take bolder steps toward integrating MSM into real-world collaboration, but they do so with higher risks in terms of generalizability and robustness.

### 8.2.4 Comparative Insights

Taken together, the three types of experimental designs explored in this thesis illustrate complementary pathways for integrating mental state monitoring into human–robot collaboration. Each comes with its own trade-offs: dual-task paradigms offer simplicity and fast implementation but limit more realistic implementation; neuropsychology-inspired tasks ensure theoretical rigor and controlled manipulations but demand incremental, time-intensive validation before reaching realistic settings; and baseline-task designs strike a balance between control and realism, though they face the challenge of task-to-task variability.

Rather than competing approaches, these designs should be viewed as a versatile toolbox. Dual-task paradigms can serve as an accessible entry

point for probing new questions, neuropsychology-inspired tasks can provide theoretically grounded bridges between laboratory control and collaboration, and baseline-task designs can push toward direct applicability in naturalistic HRC. Together, they enable researchers to balance control, realism, and feasibility depending on their goals.

By systematically developing, implementing, and critically evaluating these designs, this work not only highlights their individual strengths and limitations but also lays the groundwork for a long-term research trajectory. Future studies can build on this foundation by combining elements from these paradigms or innovating hybrid approaches, progressively moving MSM in HRC to robust, adaptive, and ecologically valid systems capable of thriving in real-world collaboration.

## 8.3 Statistical insights on subjective, behavioral, and physiological metrics

Beyond methodological design, this PhD also sought to identify empirical evidence on the reliability of different metrics of cognitive effort in HRC. Across the datasets collected, three categories of measurements were systematically evaluated. Table 8.3 synthesizes the metrics that were significantly modulated by task difficulty, and if these effects replicated across experimental sessions. Notably, the RITHMD dataset was analyzed separately by task, as distinct paradigms may elicit distinct sensitivity patterns. In contrast, for the RITHME dataset, replication across sessions could not be assessed, as only a single session per participant was available.

| Dataset | EPIIC | RITHMD | RITHMD | RITHME |
|---|---|---|---|---|
| Task | *Digit Span* | *Sternberg* | *Cog Shifting* | *Cog Shifting* |
| **Subjective** | | | | |
| ISA | .001 (.85)[1,2] | – | – | – |
| RSME | – | .001 (.86)[1,2] | .001 (.76)[1,2] | .001 (.79)[1] |
| **Accuracy** | | | | |
| % correct | .001 (.49)[1,2] | .001 (.80)[1,2] | .001 (.55)[1,2] | .001 (.60)[1] |
| F1 score | – | .001 (.90)[1,2] | .001 (.49)[1,2] | .001 (.58)[1] |
| **Response time** | | | | |
| RT | ns | .001 (.84)[1,2] | .001 (.78)[1,2] | .001 (.73)[1] |
| **EEG** | | | | |
| Frontal $\theta$ | ns | ns | ns | ns |
| Frontal $\alpha$ | ns | ns | ns | ns |
| Frontal $\beta$ | ns | ns | ns | .01 (.36)[1] |
| Frontal Engagement | ns | ns | ns | ns |
| Central $\theta$ | ns | ns | ns | ns |
| Central $\alpha$ | ns | ns | ns | ns |
| Central $\beta$ | ns | .01 (.30)[1] | .01 (.35)[1,2] | .05 (.23)[1] |
| Central Engagement | ns | .05 (.17) | .001 (.41)[1,2] | .01 (.36)[1] |
| Parietal $\theta$ | ns | ns | ns | ns |
| Parietal $\alpha$ | ns | ns | ns | ns |
| Parietal $\beta$ | ns | ns | ns | ns |
| Parietal Engagement | ns | ns | ns | ns |
| **ECG** | | | | |
| HR | – | .001 (.48)[2] | ns | .01 (.42)[1] |
| HRV | – | .001 (.49)[1,2] | .001 (.39)[1,2] | .001 (.55)[1] |
| **PPG** | | | | |
| HR | ns | .01 (.29)[1] | ns | .05 (.22)[1] |
| HRV | ns | .01 (.24)[1] | .05 (.21)[1] | ns |
| **EDA** | | | | |
| Phasic | .05 (.16)[1,2] | .001 (.36)[1,2] | ns | ns |
| Tonic | ns | ns | ns | .05 (.24)[1] |

Table 8.2: Summary of difficulty effects across datasets. Each cell indicates the $p$-value (with effect size in parentheses) and the session(s) in which the effect reached significance (RITHME only included one session). A value without a session number indicates significance when considering all data collectively but not within any individual session. "ns" denotes non-significant effects.

### 8.3.1 Subjective metrics

To assess perceived cognitive effort during the interaction, two specialized self-report tools were employed: the Instantaneous Self-Assessment (ISA) [166, 167] and the Rating Scale for Mental Effort (RSME) [199]. The widely used NASA-TLX [30] was deliberately excluded despite its popularity, as it evaluates six distinct dimensions of effort, whereas the presented studies were specifically concerned with only one of them. Selecting a specialized scale was therefore judged more appropriate than isolating a single dimension from NASA-TLX, which would have undermined the interpretability and relevance of the instrument.

In practice, both ISA and RSME were sensitive to task difficulty, showing significant effects in the experiments where they were applied. However, RSME was ultimately preferred, as its finer granularity allowed for greater variability and precision in participant responses.

Nevertheless, an important limitation of subjective metrics must be emphasized. Their robustness in reflecting the impact of difficulty levels, while valuable for verifying that manipulations were perceived as intended, paradoxically reduces their utility as predictors of behavioral or physiological responses to the same effect. In other words, although ISA and RSME provide a minimal safeguard that difficulty manipulation has been successfully induced, they cannot guarantee that this subjective perception is mirrored in more objective measures, an issue widely acknowledged in the literature [226, 227].

For mental state monitoring in HRC, this means that subjective metrics remain useful to validate task manipulations, but they are insufficient as standalone indicators and should be integrated with behavioral and/or physiological features to yield meaningful insights [52].

### 8.3.2 Behavioral metrics

Behavioral metrics such as accuracy and response time offered direct and objective indicators of task performance. Although inherently task-dependent and therefore difficult to compare across experimental designs, they were consistently sensitive to manipulations of cognitive effort in the datasets presented here, thereby supporting the validity of the experimental paradigms. Yet, this also raises a circular limitation: in these studies, behavioral measures were not emergent properties of HRC but rather artificially engineered to verify hypotheses within controlled conditions.

As such, while behavioral metrics can serve as reliable indicators for mental state monitoring in HRC, their design must be aligned with their intended

purpose. In laboratory settings, artificially introduced metrics are useful for validating hypotheses and experimental control. However, in more ecological and dynamic HRC contexts, the challenge shifts toward identifying and designing behavioral indicators that emerge naturally from the collaboration, such as fluency-related [8] or error-related [228] metrics. Although not the primary focus of this work, this issue has already been partially addressed in the literature [42, 34], and developing such indicators represents an important avenue for MSM for real-world HRC, with or without physiological signals.

### 8.3.3 Physiological metrics

Physiological metrics exhibited highly heterogeneous sensitivity to task difficulty across datasets. Among the three domains considered, cardiac and brain measures emerged as the most consistent indicators, while electrodermal features produced more variable or context-dependent effects. This variability echoes a long-standing debate in the literature : while many studies and reviews have listed physiological indicators of cognitive effort [141, 204, 229, 140], their inherent complexity and variability can not guarantee a single measure to be reliably valid for all application contexts.

For brain activity, EEG measures provided mixed outcomes across datasets. The power in the $\theta$ and $\alpha$ bands failed to reach significance in any study, which is surprising given their frequent association with cognitive effort in the literature [143]. By contrast, $\beta$ power—often reported as only moderately sensitive to workload [143]—emerged here as the most responsive, alongside the central Engagement Index [84]. Both markers were significantly modulated by task difficulty at central sites in three out of four datasets (both tasks in RITHMD and the one in RITHME), although the effect appeared in only one session of the Sternberg task. Additionally, frontal $\beta$ showed significance in the RITHME dataset. The saliency of $\beta$ activity, classically associated with sensorimotor processes and the maintenance of cognitive control [230], could be partly explained by the nature of the experiments, which placed strong demands on participants' ability to process, memorize, and recall the robot's actions.

On a different perspective, both RITHMD and RITHME relied on the cognitive shifting task but used different hardware: medical-grade wet EEG in RITHMD versus wearable dry EEG in RITHME. As expected, the dry system produced weaker effects, consistent with lower signal-to-noise ratios in mobile recordings [76]. This highlights ongoing challenges for achieving reliable wearable EEG, both at the hardware level by improving sensor quality, and at the software level by improving signal processing algorithms. Never-

theless, the fact that mobile EEG still captured relevant effects underscores its promise as a practical tool for real-world HRC.

For cardiac activity, features derived from both ECG and PPG consistently reflected task difficulty, with HRV emerging as the most reliable measure. As established in the literature, ECG-based HRV reliably decreases with increasing cognitive demand [204], and this pattern was replicated across all datasets analyzed here. By contrast, mean HR produced less consistent results: when significant, it decreased with increasing task difficulty, contradicting the well-established association of higher workload with elevated heart rate. In addition, these HR effects were limited to the RITHMD Sternberg task and the RITHME cognitive shifting task, failing to replicate across the RITHMD cognitive shifting dataset. PPG-derived measures followed the same general pattern, but with weaker effects: both HR and HRV effects were smaller in magnitude, and mean HRV did not capture task modulation in RITHME at all. These findings reinforce ECG as the gold standard for cardiac monitoring, specially the HRV metric, while also suggesting that PPG, despite its reduced sensitivity, could serve as a more practical alternative in wearable setups, provided that technological progress continues to improve its accuracy.

For electrodermal activity, results were less consistent. Phasic activity reliably tracked increases in task difficulty in both EPIIC and the RITHMD Sternberg task, with responses systematically rising alongside cognitive effort in both sessions, in line with the literature [140]. Yet it failed to replicate in the RITHMD and RITHME cognitive shifting tasks. Tonic activity reached significance only once, in RITHME and without replicating at the subgroup level. These findings suggest that while EDA phasic activity can provide meaningful insights under certain conditions, its reliability do not generalize well to all tasks.

Overall, cardiac HRV consistently emerged as the most reliable physiological marker of cognitive effort, while central $\beta$ power and central Engagement Index also showed surprising yet stable sensitivity. Although they may be partially redundant, since the Engagement Index incorporates $\beta$ power in its calculation. EDA phasic responses provided occasional complementary insights but lacked consistency across tasks. Importantly, comparisons between medical-grade and wearable devices highlighted that wearable alternatives (both dry EEG and PPG) systematically underperformed in accuracy compared to their medical-grade counterparts. While this outcome was expected, it carries a dual implication: caution is needed when interpreting wearable-derived signals, but the ability of these devices to still capture meaningful effects provides encouraging prospects for real-world HRC applications.

## 8.4 MSM insights from physiological data

A core empirical contribution of this PhD lies in the systematic benchmarking of classification pipelines for cognitive effort across multiple datasets. To guarantee fair comparison, all datasets were processed within a consistent framework, ensuring that differences in performance solely come from the data. However, it is important to note that the EPIIC dataset involved a more demanding three-class classification problem, whereas all other datasets were restricted to two-class classification. This distinction should be carefully considered when interpreting performance differences across studies.

### 8.4.1 Brain activity

| Dataset | EPIIC | RITHMD | RITHMD | RITHME |
|---|---|---|---|---|
| Task | *Digit Span* | *Sternberg* | *Cog Shifting* | *Cog Shifting* |
| **Preprocessing** | | | | |
| Minimal | .473 | .723 | .756 | .708 |
| ASR | .532 | .716 | .746 | .682 |
| ICA | .464 | .705 | .729 | .696 |
| **Dimensionality** | | | | |
| Full | .528 | .727 | .757 | .725 |
| CSP | .451 | .702 | .731 | .665 |
| **Feature** | | | | |
| beta | .536 | .765 | .795 | .752 |
| all | .494 | .747 | .772 | .71 |
| full | .461 | .755 | .776 | .683 |
| alpha | .488 | .68 | .709 | .678 |
| theta | .467 | .625 | .666 | .653 |
| **Classifier** | | | | |
| SVM | .509 | .731 | .76 | .725 |
| RF | .471 | .7 | .736 | .69 |
| LDA | .488 | .712 | .735 | .671 |

Table 8.3: Comparison of preprocessing, dimensionality reduction, feature domains, and classifiers for cognitive effort classification based on brain activity. Reported values correspond to mean classification accuracy, with the color gradient indicating performance level. Yet, EPIIC involved a 3-class classification problem, whereas all other datasets were restricted to 2-class classification.

For brain activity, multiple preprocessing strategies (None, ICA, ASR) were applied to EEG recordings and compared together. Covariance matrices were sytematically extracted as features but from various frequency range : in theta (4-8Hz), alpha (8-12Hz) and beta (12-30Hz) cannonical range, in a range encompassing all these three bands (4-30Hz) and in an extra large band encompassing full frequency power from 1 to 100Hz. The covariance matrices were subsequently projected from the Riemannian space into the Euclidean space, where they could be directly used as features for classification. Common spatial patterns (CSP) algorithms were applied and not over

these features to test the impact of dimensionality reduction. The result of these variations are reported in table 8.3.

Across most datasets, minimal preprocessing pipelines yielded slightly better performances than more complex approaches such as ASR or ICA. The only exception was the EPIIC dataset, where ASR provided an advantage, likely because robot-induced artifacts were particularly impactful in this setup.

In terms of dimensionality reduction, retaining the full covariance matrix consistently outperformed the use of CSP-based reduction. This suggests that constraining the feature space too aggressively may remove informative variance relevant for mental state decoding.

When examining the feature sets, the $\beta$ band clearly emerged as the most reliable and discriminative source of information, aligning with the statistical analyses reported in Section 8.3.3. In contrast, $\theta$ and $\alpha$ bands systematically underperformed, reflecting their lack of sensitivity to task difficulty in these experiments. Broader feature sets, such as "all" or "huge", produced good results but never exceeded the performance of the $\beta$ band alone, indicating that more information may dilute the signal-to-noise ratio instead of enhancing classification.

Finally, regarding classification algorithms, SVM consistently delivered the strongest performances across datasets, outperforming both RF and LDA.

Taken together, these results suggest that the most effective MSM pipeline for EEG on the current data combines minimal preprocessing, full covariance matrices focused on the $\beta$ band, and SVM classifier.

### 8.4.2 Cardiac activity

For cardiac activity, both HR and HRV features were extracted from ECG and PPG recordings. PCA-based aggregation was explored to reduce feature redundancy.

| Dataset | EPIIC | RITHMD | RITHMD | RITHME |
|---|---|---|---|---|
| Task | *Digit Span* | *Sternberg* | *Cog Shifting* | *Cog Shifting* |
| **Signal** | | | | |
| ECG | – | .557 | .586 | .621 |
| PPG | .394 | .539 | .561 | .599 |
| **Dimensionality** | | | | |
| PCA | .398 | .552 | .579 | .619 |
| Simple | .393 | .543 | .571 | .606 |
| **Feature** | | | | |
| HRV | .4 | .549 | .576 | .613 |
| all | .397 | .548 | .576 | .615 |
| HR | .389 | .547 | .573 | .61 |
| **Classifier** | | | | |
| RF | .396 | .554 | .578 | .619 |
| SVM | .397 | .543 | .569 | .612 |
| LDA | .39 | .548 | .573 | .6 |

Table 8.4: Comparison of sensor, dimensionality reduction, feature range, and classifiers across datasets for cognitive effort classification using cardiac activity. Values represent mean accuracy and color gradient increases with performance.

For signal type, ECG consistently outperformed PPG across datasets. While both modalities were able to capture difficulty effects, ECG systematically delivered higher classification accuracies. PPG followed the same overall trends but with slightly reduced performance, confirming it as a more wearable but less precise alternative to ECG.

Regarding dimensionality reduction, PCA provided a small but systematic advantage over simpler feature selection. This suggests that compacting cardiac features into a reduced latent space helps mitigate redundancy and noise without discarding key information.

When examining the feature sets, performances were highly consistent across variants. Feature groups including all HRV metrics, all HR metrics or both ("all") yielded all similar results. This indicates that most of the metrics explore are redundant, and classification benefits slightly from combining them rather than restricting the feature pool.

For classifiers, RF produced the best overall results, outperforming both SVM and LDA across most datasets.

Taken together, these results indicate that the most effective MSM pipeline

for cardiac activity on these data uses ECG recordings, PCA-based dimensionality reduction, all features extracted from the signal, and RF classifier.

### 8.4.3 Electrodermal activity

For electrodermal activity, both tonic and phasic components were evaluated, along with various combinations of them. PCA was again used to reduce data dimensionality.

| Dataset | EPIIC | RITHMD | RITHMD | RITHME |
|---|---|---|---|---|
| Task | *Digit Span* | *Sternberg* | *Cog Shifting* | *Cog Shifting* |
| **Dimensionality** | | | | |
| PCA | .391 | .625 | .592 | .608 |
| Simple | .396 | .594 | .578 | .591 |
| **Feature** | | | | |
| all | .396 | .614 | .598 | .614 |
| raw | .39 | .607 | .598 | .611 |
| tonic | .39 | .601 | .591 | .605 |
| phasic | .334 | .614 | .545 | .56 |
| **Classifier** | | | | |
| RF | .491 | .697 | .646 | .671 |
| SVM | .356 | .605 | .594 | .601 |
| LDA | .335 | .526 | .515 | .525 |

Table 8.5: Comparison of sensor, dimensionality reduction, feature range, and classifiers across datasets for cognitive effort classification using electrodermal activity. Values represent mean accuracy and color gradient increases with performance.

For dimensionality reduction, results showed only small differences, but PCA provided a consistent advantage over simple feature selection in three out of four datasets. This suggests that compacting EDA features into a reduced latent space helps capture the most informative components while filtering noise.

In terms of feature sets, performance was relatively stable across conditions, with all subsets. This indicates that, as fore cardiac activity, the choosen features may have been redundant.

For classifiers, RF clearly outperformed SVM and LDA across all datasets, often by a substantial margin.

In summary, the best-performing pipeline for EDA on the current data combines PCA-based dimensionality reduction, all features extracted from the signal, and RF classifier.

### 8.4.4 Summary about physiological activity

| Dataset | EPIIC | RITHMD | RITHMD | RITHME |
|---|---|---|---|---|
| Task | *Digit Span* | *Sternberg* | *Cog Shifting* | *Cog Shifting* |
| **Signal** | | | | |
| EEG | .586 | .825 | .855 | .849 |
| EDA | .512 | .712 | .656 | .676 |
| ECG | – | .556 | .59 | .617 |
| PPG | .409 | .546 | .559 | .612 |

Table 8.6: Comparison of the best pipeline for each signal across datasets for cognitive effort classification. Values represent mean accuracy.

Taken together, brain activity emerged as the most effective and consistent modality for classification, clearly outperforming cardiac and electrodermal activity, as observed in literature [231]. While cardiac activity had appeared as one of the strongest markers in the statistical analyses, its predictive power within machine learning pipelines turned out to be limited. Conversely, EDA achieved better-than-expected performances: Although its statistical sensitivity to task difficulty was inconsistent across datasets, classification models were still able to extract meaningful patterns from it in several cases.

This apparent divergence between statistical significance and classification accuracy highlights the complementary value of combining classical inferential approaches with data-driven modeling. Traditional statistical methods provide insights into the reliability and robustness of physiological markers, while machine learning can reveal new patterns in the data, even those partially buried in what is usually considered "noise" [232]. Such discrepancies remind us that information and noise are not always clearly separable, and that features overlooked by classical analyses can sometimes prove useful in specific contexts [128]. Yet, caution is essential with such "new" informative features as data-driven approaches do not guarantee generalizability. Any new physiological marker identified in this way must undergo careful replication across tasks and contexts to ensure it is not dataset-specific.

More broadly, these findings underscore the incomplete and evolving nature of our understanding of both physiological activity and machine learning

models. On the one hand, deepening our knowledge of cognitive and physiological mechanisms could help identify the most informative features to feed into classifiers, thereby improving both robustness and generalizability. On the other hand, progress in deciphering the decision processes of machine learning models could shed new light on which physiological features truly capture variations in cognitive effort, even when such features are traditionally dismissed as noise. Ultimately, fostering a closer dialogue between neuroscience and machine learning holds the potential to advance both fields simultaneously and may prove essential for developing reliable physiological monitoring tools in HRC.

## 8.5 Cognitive effort vs. error detection

This manuscript examined two central mental states for HRC: cognitive effort and error perception. While both aim to capture critical dimensions of the human experience in collaboration, each state imposes its own methodological and technical constraints on the MSM process.

Cognitive effort monitoring seeks to map ongoing physiological activity to task difficulty, typically evolving gradually over seconds or minutes [52]. Its most reliable physiological markers include spectral features of EEG and heart rate variability derived from ECG or PPG [140, 141]. These measures provide a global, continuous view of the human's cognitive availability, supporting robot adaptation in terms of task pacing, information delivery, or role distribution over longer interaction windows [53, 120].

By contrast, error perception monitoring targets fast, event-related processes where temporal precision is crucial. Detecting that an error occurred is not enough, the system must also link the detection to the precise event that caused it [214]. EEG is thus the gold standard due to its millisecond precision [66, 211]. In practice, reliable error perception monitoring could feed directly into the robot's decision-making system, enabling instantaneous corrections during collaboration as well as data collection for long-term refinement [70, 68].

Crucially, both approaches can be investigated within the several experimental design frameworks presented in this thesis, and both rely on similar processing pipelines: artifact removal, feature extraction, and supervised classification of physiological signals. In controlled laboratory contexts, they can be further complemented by behavioral and subjective measures, enriching interpretation.

Taken together, the findings and discussions of this work demonstrate that MSM in HRC is not restricted to a single mental state but instead

offers a flexible methodological framework that can be tailored to multiple states of interest. The key determinants lie at the two ends of the process: on one side, the selection of markers, and on the other, the form of robotic adaptation envisioned. Together, these factors define the temporal resolution required and, consequently, the range of physiological signals and processing strategies that can be effectively employed.

## 8.6 Perspectives

Looking ahead, the datasets and benchmarks developed in this PhD pave the way for several promising avenues to advance MSM in HRC. A high priority is the exploration of transfer learning, which is crucial to overcome inter-session, inter-subject, and inter-task variability [159, 156]. By reducing the need for extensive recalibration, transfer-based approaches could significantly increase the practical usability of MSM in real-world scenarios, where robustness and adaptability are indispensable.

Another powerful direction lies in the adoption of more sophisticated machine learning strategies. Ensemble methods and deep learning architectures [233] may provide powerful tools to capture the complex nature of physiological data, including cross-modal interactions between signals [158]. However, their integration into real-time adaptive systems will demand solutions to computational constraints and the longstanding challenge of interpretability [121]. In parallel, shifting from categorical classification toward continuous regression of subjective states could enable MSM to reflect the subtle gradations of human experience, offering richer and more dynamic insights than discrete labels alone [234].

Crucially, these advances cannot be achieved in isolation. Community-level initiatives will be essential, including the contribution of open databases [214], the development of standardized benchmark [235], and the sharing of reproducible experimental paradigms. Data collected on comparable or replicated tasks are especially valuable, as they provide the basis for meaningful benchmarking and accelerate convergence toward robust solutions.

The methodological framework introduced in this work also opens the door to exploring other mental states beyond cognitive effort and error perception. Mental fatigue, for instance, is a critical factor in long-term collaboration, shaping vigilance, decision-making, and safety [236]. Likewise, trust represents a key dimension for ensuring effective, sustainable human–robot collaboration [237]. Extending MSM to encompass such states would greatly enrich its relevance and impact.

The ultimate ambition of MSM is to enable adaptive robotic systems

that dynamically tailor their behavior to human cognitive states. Achieving this vision, however, requires first addressing the fundamental challenges of variability, generalizability, and interpretability. Once these foundations are established, adaptation itself will bring new scientific questions tied to the timing, modality, and frequency of its application that will need careful investigation.

Finally, this work underscores the importance of critically assessing the added value of physiology. In some contexts, behavioral or task-level data may provide sufficient insights into human states. In others, physiological signals offer unique, automated access to processes that remain invisible to performance measures alone. Establishing clear guidelines for identifying when physiology is truly needed will be essential to optimize resources and deploy MSM where it delivers the greatest benefits.

By systematically analyzing experimental designs, benchmarking pipelines, and extending the study to multiple mental states, this thesis aims at contributing to a scalable, generalizable, and community-driven future of MSM in HRC, bringing the field closer to truly adaptive, reliable, and human-centered collaboration.

# Peer-Reviewed Contributions

**Mathias Rihet**, Aurélie Clodic, Guillaume Sarthou, Anouk Sanamarti-Decool, Raphaëlle N. Roy (**2025**). A Set of Robotic Inductive Tasks to Monitor Human Cognitive Effort. *International Journal of Social Robotics* [submitted]

**Mathias Rihet**, Guillaume Sarthou, Aurélie Clodic, Raphaelle Roy (**2024**). Electrophysiological Measures for Human-Robot Collaboration Quality Assessment, In *Discovering the Frontiers of Human-Robot Interaction: Insights and Innovations in Collaboration, Communication, and Control*, Springer, pp.363-380.
https://doi.org/10.1007/978-3-031-66656-8_15

**Mathias Rihet**, Aurélie Clodic, Raphaelle Roy (**2024**). Robot Noise: Impact on Electrophysiological Measurements and Recommendations HRI '24: Companion of the 2024 ACM/IEEE *International Conference on Human-Robot Interaction*, pp.888-891.
https://doi.org/10.1145/3610978.3640708

**Mathias Rihet**, Kalou Cabrera Castillo, Sébastien Scannella, Tresols Juan Jesus Torre, Frédéric Dehais, Raphaëlle N. Roy (**2023**). Enhancing Human-Robot Interaction: Riemannian Classification and Clustering for Error Potential Detection and Error Occurrence Estimation, IntEr-HRI Competition for the Thirty-Third *International Joint Conference on Artificial Intelligence (IJCAI)*.
https://hal.science/hal-04832577v1

**Mathias Rihet**, Aurélie Clodic, Guillaume Sarthou, Sridath Tula, Raphaëlle N. Roy (**2022**). A controlled human-robot interaction experimental setup for physiological data acquisition, *IROS2022 Workshop: Artificial Intelligence for Social Robots Interacting with Humans in the Real World*.
https://hal.science/hal-04832578v1

# AI Statement

This dissertation benefited from the use of ChatGPT (OpenAI) as a tool to enhance linguistic clarity and reader-friendliness. The intellectual and scientific contributions are solely those of the author.

# Bibliography


[1] N. Mishraa, D. Goyal, and A. D. Sharma, "Issues in existing robotic service in restaurants and hotels," in *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, 2018, pp. 26–27. [Cited on page 13.]

[2] S. Edirisinghe, S. Satake, D. Brscic, Y. Liu, and T. Kanda, "Field trial of an autonomous shopworker robot that aims to provide friendly encouragement and exert social pressure," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 194–202. [Cited on page 13.]

[3] M. E. Foster, B. Craenen, A. Deshmukh, O. Lemon, E. Bastianelli, C. Dondrup, I. Papaioannou, A. Vanzo, J.-M. Odobez, O. Canévet *et al.*, "Mummer: Socially intelligent human-robot interaction in public spaces," *arXiv preprint arXiv:1909.06749*, 2019. [Cited on page 13.]

[4] K. Belhassein, G. Buisan, A. Clodic, and R. Alami, "Towards methodological principles for user studies in human-robot interaction," in *Test Methods and Metrics for Effective HRI in Collaborative Human-Robot Teams Workshop, ACM/IEEE International Conference on Human-Robot Interaction*, 2019. [Cited on page 14.]

[5] C. Jost, B. Le Pévédic, T. Belpaeme, C. Bethel, D. Chrysostomou, N. Crook, M. Grandgeorge, and N. Mirnig, *Human-Robot Interaction: Evaluation methods and their standardization.* Cham, Switzerland: Springer Nature, May 2020. [Cited on page 14.]

[6] G. Knoblich, S. Butterfill, and N. Sebanz, "Psychological Research on Joint Action," in *Psychology of Learning and Motivation.* Elsevier, 2011, vol. 54, pp. 59–101. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780123855275000036 [Cited on page 14.]

[7] A. Clodic, E. Pacherie, R. Alami, and R. Chatila, "Key elements for human-robot joint action," *Sociality and normativity for robots: philosophical inquiries into human-robot interactions*, pp. 159–177, 2017. [Cited on page 14.]

[8] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 209–218, 2019. [Cited on pages 14, 16, and 151.]

[9] S. Rossi, F. Ferland, and A. Tapus, "User profiling and behavioral adaptation for hri: A survey," *Pattern Recognition Letters*, vol. 99, pp. 3–12, 2017. [Cited on pages 14 and 49.]

[10] E. Coronado, T. Kiyokawa, G. A. G. Ricardez, I. G. Ramirez-Alpizar, G. Venture, and N. Yamanobe, "Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0," *Journal of Manufacturing Systems*, vol. 63, pp. 392–410, 2022. [Cited on pages 14, 15, 16, and 19.]

[11] A. Sciutti, M. Mara, V. Tagliasco, and G. Sandini, "Humanizing human-robot interaction: On the importance of mutual understanding," *IEEE Technology and Society Magazine*, vol. 37, no. 1, pp. 22–29, 2018. [Cited on page 14.]

[12] E. Montini, F. Daniele, L. Agbomemewa, M. Confalonieri, V. Cutrona, A. Bettoni, P. Rocco, and A. Ferrario, "Collaborative robotics: A survey from literature and practitioners perspectives," *Journal of Intelligent & Robotic Systems*, vol. 110, no. 3, p. 117, 2024. [Cited on page 15.]

[13] J. Arents, V. Abolins, J. Judvaitis, O. Vismanis, A. Oraby, and K. Ozols, "Human–robot collaboration trends and safety aspects: A systematic review," *Journal of Sensor and Actuator Networks*, vol. 10, no. 3, p. 48, 2021. [Cited on page 15.]

[14] S. Hopko, J. Wang, and R. Mehta, "Human factors considerations and metrics in shared space human-robot collaboration: A systematic review," *Frontiers in Robotics and AI*, vol. 9, p. 799522, 2022. [Cited on pages 15 and 19.]

[15] C. Carissoli, L. Negri, M. Bassi, F. A. Storm, and A. Delle Fave, "Mental workload and human-robot interaction in collaborative tasks: A

scoping review," *International Journal of Human–Computer Interaction*, pp. 1–20, 2023. [Cited on pages 15, 19, and 75.]

[16] L. Onnasch and C. L. Hildebrandt, "Impact of anthropomorphic robot design on trust and attention in industrial human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 1, pp. 1–24, 2021. [Cited on page 15.]

[17] S. Liao, L. Lin, Q. Chen, and H. Pei, "Why not work with anthropomorphic collaborative robots? the mediation effect of perceived intelligence and the moderation effect of self-efficacy," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 34, no. 3, pp. 241–260, 2024. [Cited on page 15.]

[18] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science robotics*, vol. 3, no. 21, p. eaat5954, 2018. [Cited on page 15.]

[19] W. Johal, "Research trends in social robots for learning," *Current Robotics Reports*, vol. 1, no. 3, pp. 75–83, 2020. [Cited on page 15.]

[20] L. Pu, W. Moyle, C. Jones, and M. Todorovic, "The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies," *The gerontologist*, vol. 59, no. 1, pp. e37–e51, 2019. [Cited on page 15.]

[21] N. L. Robinson, T. V. Cottier, and D. J. Kavanagh, "Psychosocial health interventions by social robots: systematic review of randomized controlled trials," *Journal of medical Internet research*, vol. 21, no. 5, p. e13203, 2019. [Cited on page 15.]

[22] E. Deng, B. Mutlu, M. J. Mataric *et al.*, "Embodiment in socially interactive robots," *Foundations and Trends® in Robotics*, vol. 7, no. 4, pp. 251–356, 2019. [Cited on page 15.]

[23] K. Youssef, S. Said, S. Alkork, and T. Beyrouthy, "A survey on recent advances in social robotics," *Robotics*, vol. 11, no. 4, p. 75, 2022. [Cited on page 15.]

[24] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib, "A survey of autonomous human affect detection methods for social robots engaged in natural hri," *Journal of Intelligent & Robotic Systems*, vol. 82, pp. 101–133, 2016. [Cited on page 15.]

[25] C. L. Bethel and R. R. Murphy, "Review of human studies methods in hri and recommendations," *Int. J. Soc. Robot.*, vol. 2, no. 4, pp. 347–359, 2010. [Cited on page 15.]

[26] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, pp. 71–81, 2009. [Cited on page 15.]

[27] A. Weiss and C. Bartneck, "Meta analysis of the usage of the godspeed questionnaire series," in *2015 24th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2015, pp. 381–388. [Cited on page 15.]

[28] T. Nomura, T. Suzuki, T. Kanda, and K. Kato, "Measurement of negative attitudes toward robots," *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, vol. 7, no. 3, pp. 437–454, 2006. [Cited on page 15.]

[29] N. Spatola, B. Kühnlenz, and G. Cheng, "Perception and evaluation in human–robot interaction: The human–robot interaction evaluation scale (hries)—a multicomponent approach of anthropomorphism," *International Journal of Social Robotics*, vol. 13, no. 7, pp. 1517–1539, 2021. [Cited on pages 15 and 58.]

[30] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183. [Cited on pages 16 and 150.]

[31] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988. [Cited on page 16.]

[32] E. R. Thompson, "Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas)," *Journal of cross-cultural psychology*, vol. 38, no. 2, pp. 227–242, 2007. [Cited on page 16.]

[33] L. Veling and C. McGinn, "Qualitative research in hri: A review and taxonomy," *International Journal of Social Robotics*, vol. 13, pp. 1689–1709, 2021. [Cited on page 16.]

[34] A. Mayima, A. Clodic, and R. Alami, "Towards robots able to measure in real-time the quality of interaction in hri contexts," *International Journal of Social Robotics*, vol. 14, no. 3, pp. 713–731, 2022. [Cited on pages 16 and 151.]

[35] S. Bensch, A. Jevtic, and T. Hellström, "On interaction quality in human-robot interaction," in *ICAART 2017 Proceedings of the 9th International Conference on Agents and Artificial Intelligence, vol. 1*. SciTePress, 2017, pp. 182–189. [Cited on page 16.]

[36] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, "Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 5048–5054. [Cited on page 16.]

[37] A. Belardinelli, "Gaze-based intention estimation: principles, methodologies, and applications in hri," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 3, pp. 1–30, 2024. [Cited on page 16.]

[38] R. Ishii and Y. I. Nakano, "An empirical study of eye-gaze behaviors: Towards the estimation of conversational engagement in human-agent communication," in *Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction*, 2010, pp. 33–40. [Cited on page 16.]

[39] D. Das, M. G. Rashed, Y. Kobayashi, and Y. Kuno, "Recognizing gaze pattern for human robot interaction," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 2014, pp. 142–143. [Cited on page 16.]

[40] H. Liu and L. Wang, "Gesture recognition for human-robot collaboration: A review," *International Journal of Industrial Ergonomics*, vol. 68, pp. 355–367, 2018. [Cited on page 16.]

[41] X. Zhao, A. M. Naguib, and S. Lee, "Kinect based calling gesture recognition for taking order service of elderly care robot," in *The 23rd IEEE international symposium on robot and human interactive communication*. IEEE, 2014, pp. 525–530. [Cited on page 16.]

[42] A. Castro, F. Silva, and V. Santos, "Trends of human-robot collaboration in industry contexts: Handover, learning, and metrics," *Sensors*, vol. 21, no. 12, p. 4113, 2021. [Cited on pages 16 and 151.]

[43] M. Alimardani and K. Hiraki, "Passive brain-computer interfaces for enhanced human-robot interaction," *Frontiers in Robotics and AI*, vol. 7, p. 125, 2020. [Cited on page 17.]

[44] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib, "Progress and prospects of the human–robot collaboration," *Autonomous robots*, vol. 42, pp. 957–975, 2018. [Cited on page 17.]

[45] F. Guo, M. Li, Q. Qu, and V. G. Duffy, "The effect of a humanoid robot's emotional behaviors on users' emotional responses: Evidence from pupillometry and electroencephalography measures," *International Journal of Human–Computer Interaction*, vol. 35, no. 20, pp. 1947–1959, 2019. [Cited on page 17.]

[46] A. Reuten, M. Van Dam, and M. Naber, "Pupillary responses to robotic and human emotions: the uncanny valley and media equation confirmed," *Frontiers in psychology*, vol. 9, p. 774, 2018. [Cited on page 17.]

[47] F. Dehais and H. Ayaz, "Progress and direction in neuroergonomics," in *Neuroergonomics*. Elsevier, 2019, pp. 3–7. [Cited on page 17.]

[48] R. Parasuraman and M. Rizzo, *Neuroergonomics: The brain at work*. Oxford University Press, 2006. [Cited on page 17.]

[49] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with computers*, vol. 21, no. 1-2, pp. 133–145, 2009. [Cited on pages 17 and 37.]

[50] B. Cain, "A review of the mental workload literature," *DTIC Document*, 2007. [Cited on page 18.]

[51] M. Mendl, "Performing under pressure: stress and cognitive function," *Applied animal behaviour science*, vol. 65, no. 3, pp. 221–244, 1999. [Cited on page 18.]

[52] F. Dehais, A. Lafont, R. Roy, and S. Fairclough, "A neuroergonomics approach to mental workload, engagement and human performance," *Frontiers in neuroscience*, vol. 14, p. 519228, 2020. [Cited on pages 18, 150, and 159.]

[53] Y. Liu, M. Habibnezhad, and H. Jebelli, "Brainwave-driven human-robot collaboration in construction," *Automation in Construction*, vol. 124, p. 103556, 2021. [Cited on pages 18, 43, 44, 49, 55, 70, 75, 146, and 159.]

[54] A. H. Memar and E. T. Esfahani, "Objective assessment of human workload in physical human-robot cooperation using brain monitoring," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 2, pp. 1–21, 2019. [Cited on pages 18, 43, and 44.]

[55] C. J. Lin and R. P. Lukodono, "Classification of mental workload in human-robot collaboration using machine learning based on physiological feedback," *Journal of Manufacturing Systems*, vol. 65, pp. 673–685, 2022. [Cited on pages 19 and 75.]

[56] R. Gervasi, K. Aliev, L. Mastrogiacomo, and F. Franceschini, "User experience and physiological response in human-robot collaboration: a preliminary investigation," *Journal of Intelligent & Robotic Systems*, vol. 106, no. 2, p. 36, 2022. [Cited on pages 19, 44, and 45.]

[57] A. Buerkle, H. Matharu, A. Al-Yacoub, N. Lohse, T. Bamber, and P. Ferreira, "An adaptive human sensor framework for human–robot collaboration," *The International Journal of Advanced Manufacturing Technology*, pp. 1–16, 2022. [Cited on pages 19, 43, and 44.]

[58] E. Verna, S. Puttero, G. Genta, M. Galetto *et al.*, "A novel diagnostic tool for human-centric quality monitoring in human-robot collaboration manufacturing," *Journal of Manufacturing Science and Engineering*, vol. 145, no. 12, p. 121009, 2023. [Cited on pages 19, 43, and 44.]

[59] P. A. Lasota, T. Fong, J. A. Shah *et al.*, "A survey of methods for safe human-robot interaction," *Foundations and Trends® in Robotics*, vol. 5, no. 4, pp. 261–349, 2017. [Cited on page 19.]

[60] B. Su, S. Jung, L. Lu, H. Wang, L. Qing, and X. Xu, "Exploring the impact of human-robot interaction on workers' mental stress in collaborative assembly tasks," *Applied Ergonomics*, vol. 116, p. 104224, 2024. [Cited on page 19.]

[61] Y. Zhou, T. Kornher, J. Mohnke, and M. H. Fischer, "Tactile interaction with a humanoid robot: Effects on physiology and subjective impressions," *International Journal of Social Robotics*, vol. 13, pp. 1657–1677, 2021. [Cited on page 19.]

[62] R. Agrigoroaie and A. Tapus, "Cognitive performance and physiological response analysis: Analysis of the variation of physiological parameters based on user's personality, sensory profile, and morningness–eveningness type in a human–robot interaction scenario," *International*

*Journal of Social Robotics*, vol. 12, no. 1, pp. 47–64, 2020. [Cited on pages 19, 44, and 45.]

[63] F. Dehais, E. A. Sisbot, R. Alami, and M. Causse, "Physiological and subjective evaluation of a human–robot object hand-over task," *Applied ergonomics*, vol. 42, no. 6, pp. 785–791, 2011. [Cited on pages 19, 44, and 45.]

[64] N. B. Sarter and D. D. Woods, "How in the world did we ever get into that mode? mode error and awareness in supervisory control," *Human factors*, vol. 37, no. 1, pp. 5–19, 1995. [Cited on page 20.]

[65] J. R. Wessel, C. Danielmeier, J. B. Morton, and M. Ullsperger, "Surprise and error: common neuronal architecture for the processing of errors and novelty," *Journal of Neuroscience*, vol. 32, no. 22, pp. 7528–7537, 2012. [Cited on page 20.]

[66] R. Chavarriaga, A. Sobolewski, and J. d. R. Millán, "Errare machinale est: the use of error-related potentials in brain-machine interfaces," *Frontiers in neuroscience*, vol. 8, p. 86996, 2014. [Cited on pages 20 and 159.]

[67] S. K. Ehrlich and G. Cheng, "Human-agent co-adaptation using error-related potentials," *Journal of neural engineering*, vol. 15, no. 6, p. 066014, 2018. [Cited on pages 20, 47, and 142.]

[68] A. F. Salazar-Gomez, J. DelPreto, S. Gil, F. H. Guenther, and D. Rus, "Correcting robot mistakes in real time using eeg signals," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 6570–6577. [Cited on pages 20, 46, 47, 122, 142, and 159.]

[69] J. DelPreto, A. F. Salazar-Gomez, S. Gil, R. Hasani, F. H. Guenther, and D. Rus, "Plug-and-play supervisory control using muscle and brain signals for real-time gesture and error detection," *Autonomous Robots*, vol. 44, pp. 1303–1322, 2020. [Cited on page 20.]

[70] S. K. Kim, E. A. Kirchner, A. Stefes, and F. Kirchner, "Intrinsic interactive reinforcement learning–using error-related potentials for real world human-robot interaction," *Scientific reports*, vol. 7, no. 1, pp. 1–16, 2017. [Cited on pages 20, 46, 47, 122, and 159.]

[71] R. Cooper, J. W. Osselton, and J. C. Shaw, *EEG technology*. Butterworth-Heinemann, 2014. [Cited on page 22.]

[72] E. Jungnickel, L. Gehrke, M. Klug, and K. Gramann, "Mobi—mobile brain/body imaging," *Neuroergonomics*, pp. 59–63, 2019. [Cited on pages 22 and 141.]

[73] M. Ali, A. H. Mosa, F. Al Machot, and K. Kyamakya, "Eeg-based emotion recognition approach for e-healthcare applications," in *2016 eighth international conference on ubiquitous and future networks (ICUFN)*. IEEE, 2016, pp. 946–950. [Cited on page 24.]

[74] J. N. Acharya and V. J. Acharya, "Overview of eeg montages and principles of localization," *Journal of Clinical Neurophysiology*, vol. 36, no. 5, pp. 325–329, 2019. [Cited on page 24.]

[75] G.-L. Li, J.-T. Wu, Y.-H. Xia, Q.-G. He, and H.-G. Jin, "Review of semi-dry electrodes for eeg recording," *Journal of Neural Engineering*, vol. 17, no. 5, p. 051004, 2020. [Cited on page 24.]

[76] K. E. Mathewson, T. J. Harrison, and S. A. Kizuk, "High and dry? comparing active dry eeg electrodes to active and passive wet electrodes," *Psychophysiology*, vol. 54, no. 1, pp. 74–82, 2017. [Cited on pages 24, 141, and 151.]

[77] M. Weiergräber, A. Papazoglou, K. Broich, and R. Müller, "Sampling rate, signal bandwidth and related pitfalls in eeg analysis," *Journal of neuroscience methods*, vol. 268, pp. 53–55, 2016. [Cited on page 25.]

[78] C. He, Y.-Y. Chen, C.-R. Phang, C. Stevenson, I.-P. Chen, T.-P. Jung, and L.-W. Ko, "Diversity and suitability of the state-of-the-art wearable and wireless eeg systems review," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 3830–3843, 2023. [Cited on pages 25 and 39.]

[79] F. Dehais, R. N. Roy, and S. Scannella, "Inattentional deafness to auditory alarms: Inter-individual differences, electrophysiological signature and single trial classification," *Behavioural brain research*, vol. 360, pp. 51–59, 2019. [Cited on pages 25 and 26.]

[80] M. Kutas and K. D. Federmeier, "Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp)," *Annual review of psychology*, vol. 62, no. 1, pp. 621–647, 2011. [Cited on page 25.]

[81] M. Eimer, "The face-specific n170 component reflects late stages in the structural encoding of faces," *Neuroreport*, vol. 11, no. 10, pp. 2319–2324, 2000. [Cited on page 25.]

[82] F. Dehais, A. Duprès, S. Blum, N. Drougard, S. Scannella, R. N. Roy, and F. Lotte, "Monitoring pilot's mental workload using erps and spectral power with a six-dry-electrode eeg system in real flight conditions," *Sensors*, vol. 19, no. 6, p. 1324, 2019. [Cited on page 25.]

[83] C. S. Herrmann, D. Strüber, R. F. Helfrich, and A. K. Engel, "Eeg oscillations: from correlation to causality," *International Journal of Psychophysiology*, vol. 103, pp. 12–21, 2016. [Cited on page 26.]

[84] A. T. Pope, E. H. Bogart, and D. S. Bartolome, "Biocybernetic system evaluates indices of operator engagement in automated task," *Biological psychology*, vol. 40, no. 1-2, pp. 187–195, 1995. [Cited on pages 26, 47, 59, and 151.]

[85] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American journal of physiology-heart and circulatory physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000. [Cited on page 26.]

[86] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 155–174, 2017. [Cited on page 27.]

[87] A. M. Bastos and J.-M. Schoffelen, "A tutorial review of functional connectivity analysis methods and their interpretational pitfalls," *Frontiers in systems neuroscience*, vol. 9, p. 175, 2016. [Cited on page 27.]

[88] M. Hamedi, S.-H. Salleh, and A. M. Noor, "Electroencephalographic motor imagery brain connectivity analysis for bci: a review," *Neural computation*, vol. 28, no. 6, pp. 999–1041, 2016. [Cited on page 27.]

[89] C. S. Herrmann, S. Rach, J. Vosskuhl, and D. Strüber, "Time–frequency analysis of event-related potentials: a brief tutorial," *Brain topography*, vol. 27, pp. 438–450, 2014. [Cited on page 27.]

[90] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial patterns underlying population differences in the background eeg," *Brain topography*, vol. 2, no. 4, pp. 275–284, 1990. [Cited on page 28.]

[91] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Common spatial pattern revisited by riemannian geometry," in *2010 IEEE International Workshop on Multimedia Signal Processing*. IEEE, 2010, pp. 472–476. [Cited on pages 28, 63, and 125.]

[92] C. M. Michel and T. Koenig, "Eeg microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review," *Neuroimage*, vol. 180, pp. 577–593, 2018. [Cited on page 28.]

[93] T. Koenig, S. Diezig, S. N. Kalburgi, E. Antonova, F. Artoni, L. Brechet, J. Britz, P. Croce, A. Custo, A. Damborská *et al.*, "Eeg-meta-microstates: towards a more objective use of resting-state eeg microstate findings across studies," *Brain topography*, vol. 37, no. 2, pp. 218–231, 2024. [Cited on page 28.]

[94] A. L. Goldberger, Z. D. Goldberger, and A. Shvilkin, *Goldberger's Clinical Electrocardiography-E-Book: A Simplified Approach*. Elsevier Health Sciences, 2023. [Cited on page 28.]

[95] P. Guzik and M. Malik, "Ecg by mobile technologies," *Journal of electrocardiology*, vol. 49, no. 6, pp. 894–901, 2016. [Cited on page 29.]

[96] M. Schaffarczyk, B. Rogers, R. Reer, and T. Gronwald, "Validity of the polar h10 sensor for heart rate variability analysis during resting state and incremental exercise in recreational men and women," *Sensors*, vol. 22, no. 17, p. 6536, 2022. [Cited on page 29.]

[97] S. Brage, N. Brage, P. W. Franks, U. Ekelund, and N. J. Wareham, "Reliability and validity of the combined heart rate and movement sensor actiheart," *European journal of clinical nutrition*, vol. 59, no. 4, pp. 561–570, 2005. [Cited on page 29.]

[98] Y. Liu and Y. Fu, "Research on the visual design elements of the living street from the perspective of human factors engineering," *Buildings*, vol. 14, no. 10, p. 3098, 2024. [Cited on page 29.]

[99] M. A. Almarshad, M. S. Islam, S. Al-Ahmadi, and A. S. BaHammam, "Diagnostic features and potential applications of ppg signal in healthcare: A systematic review," in *Healthcare*, vol. 10, no. 3. MDPI, 2022, p. 547. [Cited on page 29.]

[100] G. Boccignone, D. Conte, V. Cuculo, A. d'Amelio, G. Grossi, and R. Lanzarotti, "An open framework for remote-ppg methods and their

assessment," *Ieee Access*, vol. 8, pp. 216 083–216 103, 2020. [Cited on page 30.]

[101] W. Wang, A. C. Den Brinker, S. Stuijk, and G. De Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016. [Cited on page 30.]

[102] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: a review," *Medical and biological engineering and computing*, vol. 44, pp. 1031–1051, 2006. [Cited on page 31.]

[103] P. K. Jain and A. K. Tiwari, "Heart monitoring systems—a review," *Computers in biology and medicine*, vol. 54, pp. 1–13, 2014. [Cited on page 31.]

[104] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram," *International journal of cardiology*, vol. 166, no. 1, pp. 15–29, 2013. [Cited on page 31.]

[105] N. Pinheiro, R. Couceiro, J. Henriques, J. Muehlsteff, I. Quintal, L. Goncalves, and P. Carvalho, "Can ppg be used for hrv analysis?" in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 2945–2949. [Cited on page 31.]

[106] Kubios. (2025) Kubios blog. [Online]. Available: https://www.kubios.com/blog/ [Cited on page 31.]

[107] T. Pham, Z. J. Lau, S. A. Chen, and D. Makowski, "Heart rate variability in psychology: A review of hrv indices and an analysis tutorial," *Sensors*, vol. 21, no. 12, p. 3998, 2021. [Cited on page 32.]

[108] A. J. Camm, M. Malik, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, R. E. Kleiger *et al.*, "Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996. [Cited on page 32.]

[109] W.-H. Lin, D. Wu, C. Li, H. Zhang, and Y.-T. Zhang, "Comparison of heart rate variability from ppg with that from ecg," in *The International Conference on Health Informatics: ICHI 2013, Vilamoura, Portugal on 7-9 November, 2013*. Springer, 2014, pp. 213–215. [Cited on page 32.]

[110] S. K. Nayak, A. Bit, A. Dey, B. Mohapatra, and K. Pal, "A review on the nonlinear dynamical system analysis of electrocardiogram signal," *Journal of healthcare engineering*, vol. 2018, no. 1, p. 6920420, 2018. [Cited on page 32.]

[111] J. McNames and M. Aboy, "Reliability and accuracy of heart rate variability metrics versus ecg segment duration," *Medical and Biological Engineering and Computing*, vol. 44, pp. 747–756, 2006. [Cited on pages 32 and 120.]

[112] U. Nussinovitch, K. P. Elishkevitz, K. Katz, M. Nussinovitch, S. Segev, B. Volovitz, and N. Nussinovitch, "Reliability of ultra-short ecg indices for heart rate variability," *Annals of Noninvasive Electrocardiology*, vol. 16, no. 2, pp. 117–122, 2011. [Cited on page 32.]

[113] H. F. Posada-Quintero and K. H. Chon, "Innovations in electrodermal activity data collection and signal processing: A systematic review," *Sensors*, vol. 20, no. 2, p. 479, 2020. [Cited on pages 33, 34, 35, and 41.]

[114] W. Prokasy, *Electrodermal activity in psychological research*. Elsevier, 2012. [Cited on page 33.]

[115] M. Niemann, A. Prange, and D. Sonntag, "Towards a multimodal multisensory cognitive assessment framework," in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2018, pp. 24–29. [Cited on page 34.]

[116] V. Ronca, A. C. Martinez-Levy, A. Vozzi, A. Giorgi, P. Aricò, R. Capotorto, G. Borghini, F. Babiloni, and G. Di Flumeri, "Wearable technologies for electrodermal and cardiac activity measurements: a comparison between fitbit sense, empatica e4 and shimmer gsr3+," *Sensors*, vol. 23, no. 13, p. 5847, 2023. [Cited on page 34.]

[117] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, and D. Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 857–869, 2019. [Cited on page 35.]

[118] M. Spüler, T. Krumpe, C. Walter, C. Scharinger, W. Rosenstiel, and P. Gerjets, "Brain-computer interfaces for educational applications," *Informational environments: Effects of use, effective designs*, pp. 177–201, 2017. [Cited on pages 37 and 146.]

[119] R. Sitaram, T. Ros, L. Stoeckel, S. Haller, F. Scharnowski, J. Lewis-Peacock, N. Weiskopf, M. L. Blefari, M. Rana, E. Oblak *et al.*, "Closed-loop brain training: the science of neurofeedback," *Nature Reviews Neuroscience*, vol. 18, no. 2, pp. 86–100, 2017. [Cited on page 37.]

[120] R. N. Roy, N. Drougard, T. Gateau, F. Dehais, and C. P. Chanel, "How can physiological computing benefit human-robot interaction?" *Robotics*, vol. 9, no. 4, p. 100, 2020. [Cited on pages 37, 49, 71, and 159.]

[121] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018. [Cited on pages 38, 48, and 160.]

[122] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain–computer interfaces based on deep convolutional neural networks," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 10, pp. 3839–3852, 2019. [Cited on page 38.]

[123] M. A. Joadder, S. Siuly, E. Kabir, H. Wang, and Y. Zhang, "A new design of mental state classification for subject independent bci systems," *Irbm*, vol. 40, no. 5, pp. 297–305, 2019. [Cited on page 38.]

[124] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Trends in eeg-bci for daily-life: Requirements for artifact removal," *Biomedical Signal Processing and Control*, vol. 31, pp. 407–418, 2017. [Cited on page 38.]

[125] M. Rodríguez-Ugarte, E. Iáñez, M. Ortíz, and J. M. Azorín, "Personalized offline and pseudo-online bci models to detect pedaling intent," *Frontiers in neuroinformatics*, vol. 11, p. 45, 2017. [Cited on page 39.]

[126] S. Anwer, H. Li, M. F. Antwi-Afari, A. M. Mirza, M. A. Rahman, I. Mehmood, R. Guo, and A. Y. L. Wong, "Evaluation of data processing and artifact removal approaches used for physiological signals captured using wearable sensing devices during construction tasks," *Journal of Construction Engineering and Management*, vol. 150, no. 1, p. 03123008, 2024. [Cited on pages 39 and 40.]

[127] S. Hoffmann and M. Falkenstein, "The correction of eye blink artefacts in the eeg: a comparison of two prominent methods," *PloS one*, vol. 3, no. 8, p. e3004, 2008. [Cited on page 40.]

[128] M. Shahbakhti, M. Beiramvand, E. Nasiri, S. M. Far, W. Chen, J. Solé-Casals, M. Wierzchon, A. Broniec-Wójcik, P. Augustyniak, and V. Marozas, "Fusion of eeg and eye blink analysis for detection of driver fatigue," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 2037–2046, 2023. [Cited on pages 40 and 158.]

[129] A. Delorme, "Eeg is better left alone," *Scientific reports*, vol. 13, no. 1, p. 2372, 2023. [Cited on page 40.]

[130] D. Pollreisz and N. TaheriNejad, "Detection and removal of motion artifacts in ppg signals," *Mobile Networks and Applications*, vol. 27, no. 2, pp. 728–738, 2022. [Cited on page 41.]

[131] O. Rosanne, I. Albuquerque, R. Cassani, J.-F. Gagnon, S. Tremblay, and T. H. Falk, "Adaptive filtering for improved eeg-based mental workload assessment of ambulant users," *Frontiers in Neuroscience*, vol. 15, p. 611962, 2021. [Cited on page 41.]

[132] D. Berwal, C. Vandana, S. Dewan, C. Jiji, and M. S. Baghini, "Motion artifact removal in ambulatory ecg signal for heart rate variability analysis," *IEEE Sensors Journal*, vol. 19, no. 24, pp. 12 432–12 442, 2019. [Cited on page 41.]

[133] W. Chen, N. Jaques, S. Taylor, A. Sano, S. Fedor, and R. W. Picard, "Wavelet-based motion artifact removal for electrodermal activity," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 6223–6226. [Cited on page 41.]

[134] D. Seok, S. Lee, M. Kim, J. Cho, and C. Kim, "Motion artifact removal techniques for wearable eeg and ppg sensor systems," *Frontiers in Electronics*, vol. 2, p. 685513, 2021. [Cited on page 41.]

[135] A. Delorme, T. Sejnowski, and S. Makeig, "Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis," *Neuroimage*, vol. 34, no. 4, pp. 1443–1449, 2007. [Cited on pages 41 and 63.]

[136] N. Ille, P. Berg, and M. Scherg, "Artifact correction of the ongoing eeg using spatial filters based on artifact and brain signal topographies,"

*Journal of clinical neurophysiology*, vol. 19, no. 2, pp. 113–124, 2002. [Cited on page 41.]

[137] R. J. Croft and R. J. Barry, "Removal of ocular artifact from the eeg: a review," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 30, no. 1, pp. 5–19, 2000. [Cited on page 41.]

[138] M. Kaur, B. Singh, and Seema, "Comparison of different approaches for removal of baseline wander from ecg signal," in *Proceedings of the international conference & workshop on emerging trends in technology*, 2011, pp. 1290–1294. [Cited on page 41.]

[139] M. Blanco-Velasco, B. Weng, and K. E. Barner, "Ecg signal denoising and baseline wander correction based on the empirical mode decomposition," *Computers in biology and medicine*, vol. 38, no. 1, pp. 1–13, 2008. [Cited on page 41.]

[140] P. Ayres, J. Y. Lee, F. Paas, and J. J. Van Merrienboer, "The validity of physiological measures to identify differences in intrinsic cognitive load," *Frontiers in psychology*, vol. 12, p. 702538, 2021. [Cited on pages 41, 47, 120, 151, 152, and 159.]

[141] D. Tao, H. Tan, H. Wang, X. Zhang, X. Qu, and T. Zhang, "A systematic review of physiological measures of mental workload," *International journal of environmental research and public health*, vol. 16, no. 15, p. 2716, 2019. [Cited on pages 41, 151, and 159.]

[142] J. González, J. Ortega, M. Damas, P. Martín-Smith, and J. Q. Gan, "A new multi-objective wrapper method for feature selection–accuracy and stability analysis for bci," *Neurocomputing*, vol. 333, pp. 407–418, 2019. [Cited on page 42.]

[143] S. Chikhi, N. Matton, and S. Blanchet, "Eeg power spectral measures of cognitive workload: A meta-analysis," *Psychophysiology*, vol. 59, no. 6, p. e14009, 2022. [Cited on pages 42, 120, and 151.]

[144] V. Villani, B. Capelli, C. Secchi, C. Fantuzzi, and L. Sabattini, "Humans interacting with multi-robot systems: a natural affect-based approach," *Autonomous Robots*, vol. 44, pp. 601–616, 2020. [Cited on pages 43 and 44.]

[145] A. Bettoni, E. Montini, M. Righi, V. Villani, R. Tsvetanov, S. Borgia, C. Secchi, and E. Carpanzano, "Mutualistic and adaptive human-machine collaboration based on machine learning in an injection mould-

ing manufacturing line," *Procedia CIRP*, vol. 93, pp. 395–400, 2020. [Cited on pages 43 and 44.]

[146] M. Lagomarsino, M. Lorenzini, P. Balatti, E. De Momi, and A. Ajoudani, "Pick the right co-worker: Online assessment of cognitive ergonomics in human-robot collaborative assembly," *IEEE Transactions on Cognitive and Developmental Systems*, 2022. [Cited on pages 43, 44, 45, and 75.]

[147] D. Kulić and E. Croft, "Physiological and subjective responses to articulated robot motion," *Robotica*, vol. 25, no. 1, pp. 13–27, 2007. [Cited on pages 44 and 45.]

[148] T. Arai, R. Kato, and M. Fujita, "Assessment of operator stress induced by robot collaboration in assembly," *CIRP annals*, vol. 59, no. 1, pp. 5–8, 2010. [Cited on pages 44 and 45.]

[149] C. Lopes-Dias, A. I. Sburlea, and G. R. Müller-Putz, "Online asynchronous decoding of error-related potentials during the continuous control of a robot," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019. [Cited on pages 46, 47, 122, and 142.]

[150] M. L. Seghier and C. J. Price, "Interpreting and utilising intersubject variability in brain function," *Trends in cognitive sciences*, vol. 22, no. 6, pp. 517–530, 2018. [Cited on page 48.]

[151] C. Jeunet, E. Jahanpour, and F. Lotte, "Why standard brain-computer interface (bci) training protocols should be changed: an experimental study," *Journal of neural engineering*, vol. 13, no. 3, p. 036024, 2016. [Cited on pages 48 and 147.]

[152] E. Gibson, N. J. Lobaugh, S. Joordens, and A. R. McIntosh, "Eeg variability: Task-driven or subject-driven signal of interest?" *NeuroImage*, vol. 252, p. 119034, 2022. [Cited on page 48.]

[153] S. Saha and M. Baumert, "Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: a review," *Frontiers in computational neuroscience*, vol. 13, p. 87, 2020. [Cited on pages 48 and 70.]

[154] M. Spüler, T. Krumpe, C. Walter, C. Scharinger, W. Rosenstiel, and P. Gerjets, "Brain-computer interfaces for educational applications," in *Informational environments: Effects of use, effective designs*. Springer, 2017, pp. 177–201. [Cited on pages 48 and 147.]

[155] R. N. Roy, M. F. Hinss, L. Darmet, S. Ladouce, E. S. Jahanpour, B. Somon, X. Xu, N. Drougard, F. Dehais, and F. Lotte, "Retrospective on the first passive brain-computer interface competition on cross-session workload estimation," *Frontiers in Neuroergonomics*, vol. 3, p. 838342, 2022. [Cited on pages 48 and 70.]

[156] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 20–31, 2016. [Cited on pages 48, 70, and 160.]

[157] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018. [Cited on page 48.]

[158] E. Debie, R. F. Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, and H. A. Abbass, "Multimodal fusion for objective assessment of cognitive workload: A review," *IEEE transactions on cybernetics*, vol. 51, no. 3, pp. 1542–1555, 2019. [Cited on pages 48 and 160.]

[159] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in eeg signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, 2021. [Cited on pages 48, 70, and 160.]

[160] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan, "Inter-subject transfer learning with an end-to-end deep convolutional neural network for eeg-based bci," *Journal of neural engineering*, vol. 16, no. 2, p. 026007, 2019. [Cited on page 48.]

[161] R. Schulz, P. Kratzer, and M. Toussaint, "Preferred interaction styles for human-robot collaboration vary over tasks with different action types," *Frontiers in neurorobotics*, vol. 12, p. 36, 2018. [Cited on page 49.]

[162] C. Kothe, D. Medine, C. Boulay, M. Grivich, and T. Stenner, "Lab streaming layer," 2014. [Online]. Available: https://github.com/sccn/labstreaminglayer [Cited on page 55.]

[163] A. Ojeda and C. Kothe, "Extensible data format," 2015. [Online]. Available: https://github.com/sccn/xdf [Cited on page 55.]

[164] J. F. Veale, "Edinburgh handedness inventory–short form: a revised version based on confirmatory factor analysis," *Laterality: Asymmetries of Body, Brain and Cognition*, vol. 19, no. 2, pp. 164–177, 2014. [Cited on page 57.]

[165] B. A. Diaz, S. Van Der Sluis, S. Moens, J. S. Benjamins, F. Migliorati, D. Stoffers, A. Den Braber, S.-S. Poil, R. Hardstone, D. Van't Ent *et al.*, "The amsterdam resting-state questionnaire reveals multiple phenotypes of resting-state cognition," *Front. Hum. Neurosci.*, vol. 7, p. 446, 2013. [Cited on page 57.]

[166] A. J. Tattersall and P. S. Foord, "An experimental evaluation of instantaneous self-assessment as a measure of workload," *Ergonomics*, vol. 39, no. 5, pp. 740–748, 1996. [Cited on pages 58 and 150.]

[167] G. D. Flumeri, G. Borghini, P. Aricò, A. Colosimo, S. Pozzi, S. Bonelli, A. Golfetti, W. Kong, and F. Babiloni, "On the use of cognitive neurometric indexes in aeronautic and air traffic management environments," in *Int. Workshop on Symbiotic Interaction*. Springer, 2015, pp. 45–56. [Cited on pages 58 and 150.]

[168] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. S. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Front. Neurosci.*, vol. 7, no. 267, pp. 1–13, 2013. [Cited on page 59.]

[169] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. A. Chen, "Neurokit2: A python toolbox for neurophysiological signal processing," *Behavior research methods*, vol. 53, no. 4, pp. 1689–1696, 2021. [Cited on pages 59 and 60.]

[170] N. Barascud, "Meegkit," october 2021. [Online]. Available: https://github.com/nbara/python-meegkit [Cited on page 59.]

[171] A. Barachant, Q. Barthélemy, J.-R. King, A. Gramfort, S. Chevallier, P. L. C. Rodrigues, E. Olivetti, V. Goncharenko, G. W. vom Berg, G. Reguig, A. Lebeurrier, E. Bjäreholt, M. S. Yamamoto, P. Clisson, and M.-C. Corsi, "pyriemann/pyriemann: v0.3," Jul. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7547583 [Cited on pages 59 and 125.]

[172] S. Blum, N. S. Jacobsen, M. G. Bleichner, and S. Debener, "A riemannian modification of artifact subspace reconstruction for eeg arti-

fact handling," *Frontiers in human neuroscience*, vol. 13, p. 141, 2019. [Cited on page 59.]

[173] E. Combrisson and K. Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of neuroscience methods*, vol. 250, pp. 126–136, 2015. [Cited on page 64.]

[174] D. Das Chakladar and P. P. Roy, "Cognitive workload estimation using physiological measures: a review," *Cognitive Neurodynamics*, vol. 18, no. 4, pp. 1445–1465, 2024. [Cited on page 70.]

[175] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *Int J Soc Robot*, vol. 5, pp. 291–308, 2013. [Cited on page 70.]

[176] P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme, "From characterising three years of hri to methodology and reporting recommendations," in *Int Conf Hum-Robot Interaction (hri)*. IEEE, 2016, pp. 391–398. [Cited on page 70.]

[177] M. S. Prewett, R. C. Johnson, K. N. Saboe, L. R. Elliott, and M. D. Coovert, "Managing workload in human–robot interaction: A review of empirical studies," *Computers in Human Behavior*, vol. 26, no. 5, pp. 840–856, 2010. [Cited on page 75.]

[178] C. Caiazzo, M. Savkovic, M. Pusica, D. Milojevic, M. C. Leva, and M. Djapan, "Development of a neuroergonomic assessment for the evaluation of mental workload in an industrial human–robot interaction assembly task: A comparative case study," *Machines*, vol. 11, no. 11, p. 995, 2023. [Cited on page 75.]

[179] F. Nenna, V. Orso, D. Zanardi, and L. Gamberini, "The virtualization of human–robot interactions: A user-centric workload assessment," *Virtual Reality*, vol. 27, no. 2, pp. 553–571, 2023. [Cited on page 75.]

[180] C. Lim, J. A. Barragan, J. M. Farrow, J. P. Wachs, C. P. Sundaram, and D. Yu, "Physiological metrics of surgical difficulty and multi-task requirement during robotic surgery skills," *Sensors*, vol. 23, no. 9, p. 4354, 2023. [Cited on pages 75 and 146.]

[181] D. Bortot, M. Born, and K. Bengler, "Directly or on detours? how should industrial robots approximate humans?" IEEE, pp. 89–90, 2013. [Cited on pages 75 and 146.]

[182] S. Esmaeili Bijarsari, "A current view on dual-task paradigms and their limitations to capture cognitive load," *Frontiers in psychology*, vol. 12, p. 648586, 2021. [Cited on pages 75 and 146.]

[183] A. Skulmowski and G. D. Rey, "Measuring cognitive load in embodied learning settings," *Frontiers in psychology*, vol. 8, p. 1191, 2017. [Cited on pages 75 and 146.]

[184] W. K. Kirchner, "Age differences in short-term retention of rapidly changing information." *Journal of experimental psychology*, vol. 55, no. 4, p. 352, 1958. [Cited on pages 76, 144, and 146.]

[185] A. M. Owen, K. M. McMillan, A. R. Laird, and E. Bullmore, "N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies," *Human brain mapping*, vol. 25, no. 1, pp. 46–59, 2005. [Cited on page 76.]

[186] J. Palomäki, M. Kivikangas, A. Alafuzoff, T. Hakala, and C. M. Krause, "Brain oscillatory 4–35 hz eeg responses during an n-back task with complex visual stimuli," *Neuroscience letters*, vol. 516, no. 1, pp. 141–145, 2012. [Cited on pages 76, 83, and 119.]

[187] A. Soveri, J. Antfolk, L. Karlsson, B. Salo, and M. Laine, "Working memory training revisited: A multi-level meta-analysis of n-back training studies," *Psychonomic bulletin & review*, vol. 24, pp. 1077–1096, 2017. [Cited on page 76.]

[188] A.-M. Brouwer, M. A. Hogervorst, J. B. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, "Estimating workload using eeg spectral power and erps in the n-back task," *Journal of neural engineering*, vol. 9, no. 4, p. 045008, 2012. [Cited on page 76.]

[189] S. Sternberg, "High-speed scanning in human memory," *science*, vol. 153, no. 3736, pp. 652–654, 1966. [Cited on pages 76, 144, and 146.]

[190] O. Jensen and C. D. Tesche, "Frontal theta activity in humans increases with memory load in a working memory task," *European journal of Neuroscience*, vol. 15, no. 8, pp. 1395–1399, 2002. [Cited on page 76.]

[191] C. L. Baldwin and B. Penaranda, "Adaptive training using an artificial neural network and eeg metrics for within-and cross-task workload classification," *NeuroImage*, vol. 59, no. 1, pp. 48–56, 2012. [Cited on page 76.]

[192] R. N. Roy, S. Charbonnier, A. Campagne, and S. Bonnet, "Efficient mental workload estimation using task-independent eeg features," *Journal of neural engineering*, vol. 13, no. 2, p. 026019, 2016. [Cited on page 76.]

[193] M. Altamura, B. Elvevåg, G. Blasi, A. Bertolino, J. H. Callicott, D. R. Weinberger, V. S. Mattay, and T. E. Goldberg, "Dissociating the effects of sternberg working memory demands in prefrontal cortex," *Psychiatry Research: Neuroimaging*, vol. 154, no. 2, pp. 103–114, 2007. [Cited on pages 76, 83, and 119.]

[194] E. De Lissnyder, E. H. Koster, N. Derakshan, and R. De Raedt, "The association between depressive symptoms and executive control impairments in response to emotional and non-emotional information," *Cognition and Emotion*, vol. 24, no. 2, pp. 264–280, 2010. [Cited on page 76.]

[195] U. Mayr and S. W. Keele, "Changing internal constraints on action: the role of backward inhibition." *Journal of Experimental Psychology: General*, vol. 129, no. 1, p. 4, 2000. [Cited on page 76.]

[196] A. J. Whitmer and M. T. Banich, "Inhibition versus switching deficits in different forms of rumination," *Psychological science*, vol. 18, no. 6, pp. 546–553, 2007. [Cited on page 76.]

[197] W. C. Stokoe Jr, "Sign language structure: An outline of the visual communication systems of the american deaf," *Journal of deaf studies and deaf education*, vol. 10, no. 1, pp. 3–37, 2005. [Cited on page 77.]

[198] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "Psychopy2: Experiments in behavior made easy," *Behavior research methods*, vol. 51, pp. 195–203, 2019. [Cited on page 80.]

[199] F. Zijlstra and L. Van Doorn, *The construction of a scale to measure perceived effort*. University of Technology, 1985. [Cited on pages 82, 93, 110, and 150.]

[200] A. Widyanti, A. Johnson, and D. de Waard, "Adaptation of the rating scale mental effort (rsme) for use in indonesia," *International Journal of Industrial Ergonomics*, vol. 43, no. 1, pp. 70–76, 2013. [Cited on pages 82 and 93.]

[201] R. Vallat, "Pingouin: statistics in python." *J. Open Source Softw.*, vol. 3, no. 31, p. 1026, 2018. [Cited on pages 85, 96, 112, and 138.]

[202] J. L. Szalma, J. S. Warm, G. Matthews, W. N. Dember, E. M. Weiler, A. Meier, and F. T. Eggemeier, "Effects of sensory modality and task duration on performance, workload, and stress in sustained attention," *Human factors*, vol. 46, no. 2, pp. 219–233, 2004. [Cited on page 119.]

[203] K. Yu, I. Prasad, H. Mir, N. Thakor, and H. Al-Nashash, "Cognitive workload modulation through degraded visual stimuli: A single-trial eeg study," *Journal of neural engineering*, vol. 12, no. 4, p. 046020, 2015. [Cited on page 119.]

[204] L. Longo, C. D. Wickens, G. Hancock, and P. A. Hancock, "Human mental workload: A survey and a novel inclusive definition," *Frontiers in psychology*, vol. 13, p. 883321, 2022. [Cited on pages 119, 120, 151, and 152.]

[205] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," *International Journal of Human-Computer Studies*, vol. 77, pp. 23–37, 2015. [Cited on page 120.]

[206] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein, "Erp components on reaction errors and their functional significance: a tutorial," *Biological psychology*, vol. 51, no. 2-3, pp. 87–107, 2000. [Cited on page 122.]

[207] R. Chavarriaga, A. Sobolewski, and J. d. R. Millán, "Errare machinale est: the use of error-related potentials in brain-machine interfaces," *Frontiers in neuroscience*, vol. 8, p. 208, 2014. [Cited on pages 122 and 141.]

[208] X. Wang, H.-T. Chen, Y.-K. Wang, and C.-T. Lin, "Implicit robot control using error-related potential-based brain–computer interface," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 1, pp. 198–209, 2022. [Cited on page 122.]

[209] A. Xavier Fidêncio, C. Klaes, and I. Iossifidis, "Error-related potentials in reinforcement learning-based brain-machine interfaces," *Frontiers in human neuroscience*, vol. 16, p. 806517, 2022. [Cited on page 122.]

[210] F. Iwane, I. Iturrate, R. Chavarriaga, and J. del R Millán, "Invariability of eeg error-related potentials during continuous feedback protocols

elicited by erroneous actions at predicted or unpredicted states," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046044, 2021. [Cited on pages 122 and 141.]

[211] M. Spüler and C. Niethammer, "Error-related potentials during continuous feedback: using eeg to detect errors of different type and severity," *Frontiers in human neuroscience*, vol. 9, p. 155, 2015. [Cited on pages 123, 141, and 159.]

[212] S. K. Ehrlich and G. Cheng, "A feasibility study for validating robot actions using eeg-based error-related potentials," *International Journal of Social Robotics*, vol. 11, pp. 271–283, 2019. [Cited on page 123.]

[213] F. Iwane, A. Sobolewski, R. Chavarriaga, and J. d. R. Millán, "Eeg error-related potentials encode magnitude of errors and individual perceptual thresholds," *Iscience*, vol. 26, no. 9, 2023. [Cited on page 123.]

[214] N. Kueper, K. Chari, J. Bütefür, J. Habenicht, S. K. Kim, T. Rossol, M. Tabie, F. Kirchner, and E. A. Kirchner, "Eeg and emg dataset for the detection of errors introduced by an active orthosis device," 2023. [Cited on pages 123, 159, and 160.]

[215] N. Yeung, M. M. Botvinick, and J. D. Cohen, "The neural basis of error detection: conflict monitoring and the error-related negativity." *Psychological review*, vol. 111, no. 4, p. 931, 2004. [Cited on page 124.]

[216] B. Keysar, D. J. Barr, J. A. Balin, and J. S. Brauner, "Taking perspective in conversation: The role of mutual knowledge in comprehension," *Psychological science*, vol. 11, no. 1, pp. 32–38, 2000. [Cited on pages 127, 128, and 144.]

[217] V. Pile, S. P. Haller, C. F. Hiu, and J. Y. Lau, "Young people with higher social anxiety are less likely to adopt the perspective of another: Data from the director task," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 55, pp. 41–48, 2017. [Cited on page 128.]

[218] L. Malkin, K. Abbot-Smith, and D. Williams, "Is verbal reference impaired in autism spectrum disorder? a systematic review," *Autism & Developmental Language Impairments*, vol. 3, p. 2396941518763166, 2018. [Cited on page 128.]

[219] G. Sarthou, A. Mayima, G. Buisan, K. Belhassein, and A. Clodic, "The director task: a psychology-inspired task to assess cognitive and

interactive robot architectures," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 770–777. [Cited on page 128.]

[220] P. W. Ferrez and J. d. R. Millán, "Error-related eeg potentials generated during simulated brain–computer interaction," *IEEE transactions on biomedical engineering*, vol. 55, no. 3, pp. 923–929, 2008. [Cited on page 134.]

[221] Y. Pan, T. O. Zander, and M. Klug, "Advancing passive bcis: a feasibility study of two temporal derivative features and effect size-based feature selection in continuous online eeg-based machine error detection," *Frontiers in Neuroergonomics*, vol. 5, p. 1346791, 2024. [Cited on page 141.]

[222] Z. Ren, X. Xia, Y. Tang, B. Zhao, C. P. Wong, and D. Xiao, "Asynchronous detection of erroneous behaviors in human-robot interaction with eeg: A comparative analysis of machine learning models," *bioRxiv*, pp. 2023–09, 2023. [Cited on page 141.]

[223] H. Gunes, F. Broz, C. S. Crawford, A. R.-v. der Pütten, M. Strait, and L. Riek, "Reproducibility in human-robot interaction: Furthering the science of hri," *Current Robotics Reports*, vol. 3, no. 4, pp. 281–292, 2022. [Cited on page 144.]

[224] B. Leichtmann, V. Nitsch, and M. Mara, "Crisis ahead? why human-robot interaction user studies may have replicability problems and directions for improvement," *Frontiers in Robotics and AI*, vol. 9, p. 838116, 2022. [Cited on page 144.]

[225] F. Dehais, A. Duprès, S. Blum, N. Drougard, S. Scannella, R. N. Roy, and F. Lotte, "Monitoring pilot's mental workload using erps and spectral power with a six-dry-electrode eeg system in real flight conditions," *Sensors*, vol. 19, no. 6, p. 1324, 2019. [Cited on page 146.]

[226] Y.-Y. Yeh and C. D. Wickens, "Dissociation of performance and subjective measures of workload," *Human factors*, vol. 30, no. 1, pp. 111–120, 1988. [Cited on page 150.]

[227] S. Marino, K. Meador, D. Loring, M. Okun, H. Fernandez, A. Fessler, R. Kustra, J. Miller, P. G. Ray, A. Roy *et al.*, "Subjective perception of cognition is related to mood and not performance," *Epilepsy & Behavior*, vol. 14, no. 3, pp. 459–464, 2009. [Cited on page 150.]

[228] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in psychology*, vol. 9, p. 861, 2018. [Cited on page 151.]

[229] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," *Applied ergonomics*, vol. 74, pp. 221–232, 2019. [Cited on page 151.]

[230] A. K. Engel and P. Fries, "Beta-band oscillations—signalling the status quo?" *Current opinion in neurobiology*, vol. 20, no. 2, pp. 156–165, 2010. [Cited on page 151.]

[231] M. A. Hogervorst, A.-M. Brouwer, and J. B. Van Erp, "Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload," *Frontiers in neuroscience*, vol. 8, p. 322, 2014. [Cited on page 158.]

[232] Z. Khademi, F. Ebrahimi, and H. M. Kordy, "A review of critical challenges in mi-bci: From conventional to deep learning methods," *Journal of Neuroscience Methods*, vol. 383, p. 109736, 2023. [Cited on page 158.]

[233] B. Rim, N.-J. Sung, S. Min, and M. Hong, "Deep learning in physiological signal data: A survey," *Sensors*, vol. 20, no. 4, p. 969, 2020. [Cited on page 160.]

[234] L. D. Fiederer, M. Völker, R. T. Schirrmeister, W. Burgard, J. Boedecker, and T. Ball, "Hybrid brain-computer-interfacing for human-compliant robots: inferring continuous subjective ratings with deep regression," *Frontiers in Neurorobotics*, vol. 13, p. 76, 2019. [Cited on page 160.]

[235] S. Chevallier, I. Carrara, B. Aristimunha, P. Guetschel, S. Sedlar, B. Lopes, S. Velut, S. Khazem, and T. Moreau, "The largest eeg-based bci reproducibility study for open science: the moabb benchmark," *arXiv preprint arXiv:2404.15319*, 2024. [Cited on page 160.]

[236] R. N. Roy, S. Bonnet, S. Charbonnier, and A. Campagne, "Mental fatigue and working memory load estimation: interaction and implications for eeg-based passive bci," in *2013 35th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 6607–6610. [Cited on page 160.]

[237] W.-L. Hu, K. Akash, N. Jain, and T. Reid, "Real-time sensing of trust in human-machine interactions," *IFAC-PapersOnLine*, vol. 49, no. 32, pp. 48–53, 2016. [Cited on page 160.]