

Towards real-world applicability of Neuroadaptive Technologies - Investigating subject-independence, task-independence and versatility of passive brain-computer interfaces

*Von der Fakultät 1 - MINT - Mathematik, Informatik, Physik, Elektro-und
Informationstechnik der
Brandenburgischen Technischen Universität Cottbus–Senftenberg
genehmigte Dissertation zur Erlangung des akademischen Grades eines*

Doktor der Naturwissenschaften

– Dr. rer. nat. –

vorgelegt von

LENA M. ANDREEßEN

geboren am 01.07.1988 in Hamburg

Vorsitzender: Prof. Dr. Douglas W. Cunningham

Gutachter: Prof. Dr. Thorsten O. Zander

Gutachter: Dr. Fabien Lotte

Tag der mündlichen Prüfung: 29.09.2023

DOI: <https://doi.org/10.26127/BTUOpen-6652>

Teile dieser Arbeit wurden bereits veröffentlicht:

Parts of this work have already been published:

- **Andreessen, L. M., Gerjets, P., Meurers, D., and Zander, T. O. (2021).** Toward neuroadaptive support technologies for improving digital reading: a passive BCI-based assessment of mental workload imposed by text difficulty and presentation speed during reading. *User Modeling and User-Adapted Interaction*, 31(1), 75-104.

Abstract

Brain-computer interfacing (BCI) is a relatively new field of study that involves engineering, neuroscience, psychology, and physiology. A BCI is a system that allows direct communication between the brain and the environment, using signals from a brain sensing technique, e.g. electroencephalography (EEG), to identify a user's intentions and mental states in the moment they occur. This can be used for human-machine interaction (HMI) to adapt software or control a device, automatically. *Passive* BCIs are the most promising for the general population, as they can detect and interpret a user's mental state without requiring their attention. Specifically, passive BCIs can enable machines to support solving the task at hand better, by integrating the user's task-relevant perception and interpretation into its own actions. This can be used for neuroadaptive technology (NAT), which allows a machine to learn over time how the user perceives and interprets the world. However, there are still obstacles to be overcome before BCI technology can be applied in real-world situations for NAT.

In the first part of this thesis I address the issue of *subject dependence* in training data collection for classifier calibration, which can be time-consuming and impractical. I investigated the training of a *subject-independent* predictive model that is trained on a group of other users' data and applied to online-testing data of a new person. This was done in the context of training a classifier for error detection and correction. Results showed that a classifier model can be trained without user-specific calibration and with high accuracy. The number of electrodes used in training the model was also reduced. Further it was validated that the trained classifier models were based on cortical sources and not other modalities.

In a second study I address the issue of *task-dependence*. Usually, a specific calibration paradigm is generated for every specific task where a BCI is trained for online application. Some tasks potentially are based on similar cognitive processes of mental states though, making it unnecessarily effortful to design a specific calibration paradigm for each task in which the mental state can occur. Therefore I tested the application of a potentially *task-independent* calibration paradigm for mental workload assessment in a new task. This new task was a speed reading context, where subjects read texts of varying difficulty and speed. The study found that the mental workload prediction model was accurate in classifying mental workload in different reading tasks, indicating that it can be used as a task-independent classifier.

In a third study I examine if it is possible to measure neural correlates of human *moral assessment* using a passive BCI on a single-trial basis. A calibration paradigm was developed using pictures ranked as morally unacceptable and morally neutral. However, the results showed low classification accuracies and it was not possible to

reliably distinguish between a user's subjective moral evaluations on a single-trial basis using current classification approaches.

The results presented in this thesis provide solutions towards real-world applicability of NAT enabled by passive BCIs, as examples of a subject-independent and a task-independent classifier are demonstrated and discussed. Further, approaches for increased versatility of passive BCI technology are presented, that could make passive BCIs more feasible for use in real-world human-machine interaction settings in the future.

Keywords: *Passive Brain-Computer Interfaces, Mental State Assessment, Neuroadaptive Technology, Human-Machine Interaction, Electroencephalography, User Modelling, Subject Independence, Task Independence, Error Detection, Mental Workload, Autonomous Moral Agents, Value Alignment.*

Zusammenfassung

Die Schnittstelle zwischen Gehirn und Computer (BCI) ist ein relativ neues Forschungsgebiet, welches u.a. Technik, Neurowissenschaften, Psychologie und Physiologie umfasst. Ein BCI ist ein System, das eine direkte Kommunikation zwischen dem Gehirn und der Umwelt ermöglicht, indem es Signale, die durch eine Messmethode für Gehirndaten, z.B. der Elektroenzephalographie (EEG), erfasst werden, interpretiert, um die Absichten und mentalen Zustände des Benutzers zu dem Zeitpunkt an dem sie auftreten, zu erkennen. Dies kann für die Mensch-Maschine-Interaktion (HMI) genutzt werden, um automatisch Software anzupassen oder ein Gerät zu steuern. *Passive* BCIs sind für die Allgemeinbevölkerung am vielversprechendsten, da sie den mentalen Zustand eines Benutzers erkennen und interpretieren können, ohne dessen Aufmerksamkeit zu erfordern. Insbesondere können passive BCIs Maschinen in die Lage versetzen, die Bewältigung einer Aufgabe besser zu unterstützen, indem sie die aufgabenrelevante Wahrnehmung und Interpretation des Benutzers in ihre eigenen Aktionen integrieren. Dies kann für die neuroadaptive Technologie (NAT) genutzt werden, die es einer Maschine ermöglicht mit der Zeit zu lernen, wie der Nutzer die Welt wahrnimmt und interpretiert. Allerdings gibt es Hindernisse, die überwunden werden müssen, bevor die BCI-Technologie in realen Situationen für NAT eingesetzt werden kann.

Im ersten Teil dieser Arbeit befaße ich mich zunächst mit dem Problem der *Subjektabhängigkeit* bei der Sammlung von Trainingsdaten für die Kalibrierung von Klassifikatoren, die zeitaufwändig und unpraktisch sein kann. Dafür untersuchte ich das Training eines *subjektunabhängigen* prädiktiven Modells, das mit den Daten von einer Gruppe anderer Nutzer trainiert und auf die Online-Testdaten einer neuen Person angewendet wird. Diese Analyse wurde im Kontext des Trainings eines Klassifikators zur Fehlererkennung und Korrektur durchgeführt. Die Ergebnisse zeigten, dass ein Klassifikatormodell ohne benutzerspezifische Kalibrierung und mit hoher Genauigkeit trainiert werden kann. Die Anzahl der Elektroden, die für das Training des Modells genutzt wurden, wurde ebenfalls reduziert. Außerdem wurde validiert, dass die trainierten Klassifikatormodelle auf kortikalen Quellen und nicht auf anderen Modalitäten basierten.

In einer zweiten Studie habe ich die Frage der *Aufgabenabhängigkeit* untersucht. Normalerweise wird für jede spezifische Aufgabe, für die ein BCI für die Online-Anwendung trainiert wird, ein eigenes Kalibrierungsparadigma erstellt. Einige Aufgaben basieren jedoch potenziell auf ähnlichen kognitiven Prozessen mentaler Zustände, so dass es unnötig aufwändig ist, für jede Aufgabe, bei der dieser mentale Zustand auftreten kann, ein eigenes Kalibrierungsparadigma zu entwickeln. Daher habe ich die Anwendung eines potentiell *aufgabenunabhängigen* Paradigmas zur Kalibrierung eines Klassifikatormodells zur Messung der mentalen Arbeits-

belastung in einer neuen Aufgabe getestet. Bei dieser neuen Aufgabe handelte es sich um ein Schnelleseverfahren, bei dem die Probanden Texte unterschiedlicher Schwierigkeit und Geschwindigkeit lesen. Die Studie ergab, dass das Modell zur Vorhersage der mentalen Arbeitsbelastung bei verschiedenen Leseaufgaben genau klassifizieren konnte, was darauf hindeutet, dass es als aufgabenunabhängiger Klassifikator verwendet werden kann.

In einer dritten Studie habe ich untersucht, ob es möglich ist, neuronale Korrelate der menschlichen *moralischen Bewertung* mit Hilfe eines passiven BCI auf Einzeldurchgangsbasis zu messen. Ein Kalibrierungsparadigma wurde mit Bildern entwickelt, die als moralisch inakzeptabel und moralisch neutral eingestuft wurden. Die Ergebnisse zeigten jedoch eine geringe Klassifizierungsgenauigkeit und es war nicht möglich zwischen den subjektiven moralischen Einschätzungen eines Nutzers auf Einzeldurchgangsbasis mithilfe von derzeitigen Klassifizierungsansätzen zuverlässig zu unterscheiden.

Die in dieser Arbeit vorgestellten Ergebnisse bieten Lösungen für die praktische Anwendbarkeit von NAT, die durch passive BCIs ermöglicht wird, da Beispiele für einen subjektunabhängigen und einen aufgabenunabhängigen Klassifikator gezeigt und diskutiert werden. Darüber hinaus werden Ansätze für eine erhöhte Einsatzflexibilität der passiven BCI-Technologie vorgestellt, die deren zukünftige Anwendbarkeit in der realen Welt für die Mensch-Maschine-Interaktion verbessern könnten.

*In the future, our thoughts will be free
To roam and explore, as far as the sea
Passive brain-computer interfaces will allow
Us to connect and communicate, without a sound*

*No need for words or hand gestures
Our thoughts will be the only measures
We'll control devices with just a thought
Our minds will be what machines are taught*

*But as we merge with technology
We must be careful, it can be
A double-edged sword, this power we wield
We must be mindful, for it can be used for ill or for good*

ChatGPT Dec 15 Version, OpenAI (2022)

<https://chat.openai.com/chat> (accessed 16-12-22)

Prompt: "Write three verses of a ballad about passive brain-computer interfaces."

*Thoughts flow like a stream,
Non-invasive BCIs pick up the flow,
Connecting us all.*

ChatGPT Dec 15 Version, OpenAI (2022)

<https://chat.openai.com/chat> (accessed 16-12-22)

Prompt: "Write a haiku about passive brain computer interfaces."

Acknowledgements

I would like to take this opportunity to thank all those who have accompanied me on my path to my doctorate - the preparation and completion of this dissertation would not have been possible without the support and collaboration of many wonderful and talented people surrounding me.

First and foremost, I would like to thank my doctoral supervisor Prof. Dr. Thorsten O. Zander for getting me interested in the field of passive Brain-Computer Interfaces and Neuroadaptive Technologies through his teaching and for giving me the opportunity to work scientifically in his research group at an early stage of my academic path. This laid the foundations for the work that is now presented here. In particular, I would like to express my heartfelt gratitude for the stimulating scientific discussions and impulses of these past years, as well as for the intensive support, personal commitment, creative ideas, constructive criticism and constant motivation during the preparation of my dissertation and in other research projects.

I would also like to thank Dr. Fabien Lotte for taking the time and effort to read, reflect on and review my dissertation in great detail. I would also like to express my gratitude to him for the constructive and helpful advice and the exciting and motivating scientific discussions over the last few years.

I would like to express my deepest appreciation to my committee, Prof. Dr. Erich Schneider and Prof. Dr. Stefan Glasauer, for taking the time to read and review my dissertation and challenge it in the defense. I further would like to thank Prof. Dr. Douglas W. Cunningham for his role as the chairman of the doctoral commission.

I also want to specifically thank all the staff at the Chair for Neuroadaptive Human-Computer Interaction at BTU Cottbus-Senftenberg and the members of Team PhyPA over the last few years for their daily support, the discussions about exciting projects and all the wonderful and great adventures, experiences and food we have shared as a team, as well as the challenges we conquered together. Here, my special thanks go to Juliane Pawlitzki and Dr. Laurens Krol.

Furthermore, I would like to thank Dr. Marius Klug for his expertise and advice on ICA and scientific discussions, as well as Prof. Dr. Klaus Gramann and the researchers of the Department of Biological Psychology and Neuroergonomics at TU Berlin, where I enjoyed working and where I could always discuss new ideas and problems with great and motivated researchers who helped me find new inspiration and solutions to problems.

Additionally, I would like to thank Prof. Dr. Hartmut Leuthold and Prof. Dr. Jos van Berkum again for sharing their datasets with me, so I could analyse them for the third study presented in this dissertation.

I would like to additionally thank the participants who took the time to take part in my different studies and pre-studies - without you none of this research would have been possible.

Finally, I want to express how grateful I am to my friends and family for all their time and support, without which I would never have been able to complete this work. Therefore, I would like to thank Juliane, Lara, Steffen, Christopher, Sina, Neele and Mette for their encouragement during the preparation of my defense, as well as for always cheering me on with your friendship and support. I further want to thank my parents for their unconditional love, trust and support throughout my life, without which I would probably not be who I am today. Finally, I would like to thank my husband for his encouragement, humor, unconditional support and putting up with many stressful moments during the completion of this work - I am sure it was not always easy. I would also like to thank my son for accompanying me in this world for the rest of the completion of my doctorate and for motivating and enchanting me with his sweet daily laughter, his curiosity and incredible development, that already went much faster than my writing of this dissertation.

CONTENTS

ABSTRACT	i
ZUSAMMENFASSUNG	iii
ACKNOWLEDGEMENTS	vii
LIST OF TERMS AND ABBREVIATIONS	xiii
1 Introduction	1
1.1 Motivation	1
1.2 Scope and structure of this thesis	2
2 State of the art: Brain-Computer Interfaces (BCI)	5
2.1 Preliminary note	5
2.2 Definition and components of BCIs	5
2.3 Signal acquisition	8
2.4 Electroencephalography (EEG)	9
2.4.1 Event-related Potentials (ERP)	11
2.4.2 Spectral changes in neural activity	13
2.4.3 Independent Component Analysis (ICA)	15
2.5 Signal processing and machine learning steps in a BCI	17
2.5.1 Pre-Processing	18
2.5.2 Feature extraction	19
2.5.3 Classification	21
2.5.4 Model application and validation of BCI approaches	24
2.6 Passive BCIs	26
2.6.1 Categorisation into active, reactive and passive BCIs	26
2.6.2 Passive BCIs for improved Human-Machine Interaction (HMI) .	27
2.6.3 Neuroadaptive Technology (NAT) enabled through passive BCIs	37
2.7 Challenges for passive BCI applications in real-world contexts	38
2.7.1 Sensor technology	38
2.7.2 Computation and classifier performance	39
2.7.3 Artefacts and confounds	40
2.7.4 Legal, security and ethical considerations	41
2.7.5 Subject-dependence	42
2.7.6 Task-dependence	45
2.7.7 Novel useful neural correlates for increased versatility in pas- sive BCI applications	47
3 Study 1: Subject- and session independent single-trial classification of automation error perception	49

3.1	Motivation and summary of the study	49
3.1.1	Experimental paradigm mimicking error-prone HMI and expected ErrP	50
3.1.2	Investigating subject- and session independence in an error-detection passive BCI	50
3.2	Study parts 1 to 7: Methods, results and discussion	55
3.2.1	Part 1: Neurophysiology of a subject-specific error-detection classifier	55
3.2.2	Part 2: Calibrating subject-specific error-classifiers on data cleaned from eye activity	71
3.2.3	Part 3: Session independence - Applying subject-specific classifier models from a first recording session to new subject-specific testdata from a second recording session	81
3.2.4	Part 4: Session independence - Training a group classifier on calibration data from the first recording session and applying it to new subject-specific testdata from a second recording session	84
3.2.5	Part 5: Subject-independence - Applying subject-specific and group classifier models to subject-specific testdata from a different subject	91
3.2.6	Part 6: Reducing the number of EEG channels used and omitting channels recording eye activity	96
3.2.7	Part 7: Session- and Subject independence - Application of the group classifier to data from new subjects from a third recording session	106
3.3	General discussion	119
3.4	Conclusion	123
4	Study 2: Task-independent mental workload assessment in a speed-reading context	127
4.1	Summary of the study	127
4.2	Introduction and motivation	128
4.2.1	Task-independent classification of spectral changes through mental workload	129
4.2.2	Mental workload and text comprehension	130
4.2.3	Rapid serial visual presentation (RSVP) applications for variable text presentation speed	132
4.3	Methods	134
4.3.1	Subjects	134
4.3.2	Apparatus	134
4.3.3	Stimuli and procedure	134
4.4	Data Analysis	138
4.4.1	EEG data	138
4.4.2	RTLX	142
4.4.3	Literal text comprehension	142
4.5	Results	144
4.5.1	EEG data	144

4.5.2	RTLX	149
4.5.3	Literal text comprehension	149
4.6	Discussion	151
4.7	Conclusion	156
5	Study 3: Monitoring neural correlates of human moral assessments for future human-machine value alignment	159
5.1	Motivation and summary of the study	159
5.2	Theoretical background	160
5.2.1	Autonomous Systems	160
5.2.2	Artificial Moral Agents	162
5.2.3	Machine Morality for Human-Machine Interaction	163
5.2.4	Approaches for Human-Machine Value Alignment	164
5.2.5	Neurophysiology of Human Morality	165
5.2.6	Utilizing Passive BCIs for Neuroadaptive Human-Machine Value Alignment	166
5.3	Research questions	168
5.4	Methods	169
5.4.1	Participants	169
5.4.2	Apparatus	169
5.4.3	Stimuli	169
5.4.4	Procedure	170
5.5	Data Analysis	172
5.5.1	ERP analysis	172
5.5.2	Classification on stimuli features	172
5.6	Results	173
5.6.1	ERP analyses	173
5.6.2	Classification analyses	176
5.7	Discussion	181
5.7.1	General discussion of the results	181
5.7.2	Limitations and restrictions	182
5.7.3	Outlook and conclusion	183
6	Conclusions	185
6.1	General conclusions	185
6.2	Future research and perspectives	188
	REFERENCES	193
	LIST OF FIGURES	217
	LIST OF TABLES	221

Appendices

List of terms and abbreviations

ACC Anterior cingulate cortex
AI Artificial intelligence
ALS Amyotrophic lateral sclerosis
AMA Autonomous moral agent
ANOVA Analysis of variance
BCI Brain-Computer Interface
CAR Common average reference
CSP Common spatial patterns
DNN Deep neural network
ECI Eye-computer interface
ECoG Electrocorticography
EEG Electroencephalography
EOG Electrooculogram
ER Error rate
ERD Event-related desynchronisation
ERN Error-related negativity
ERP Event-related potential
fbCSP Filter bank common spatial patterns
FBTSP Filter Bank Tangent Space Classifier
FFT Fast fourier transformation
fMRI Functional magnetic resonance imaging
HMI Human-machine interaction
IAPS International Affective Picture System
IC Independent component
ICA Independent Component Analysis
ISI Inter stimulus interval
LDA Linear discriminant analysis
LPP Late positive potential
MEG Magnetic encephalography
MNI Montreal Neurological Institute
NAT Neuroadaptive technology
NIRS Near-infrared spectroscopy
ORP Optimal recognition point
pBCI Passive brain-computer interface

Pe Error positivity
PSD Power spectral density
RLR-game Rotation-Left-Right game
RSVP Rapid serial visual presentation
RTLX Raw task load index questionnaire
SD Standard deviation
SMID Socio-moral image database
SNR Signal-to-noise ratio
SSVEP Steady-state visually evoked potentials
VR Virtual reality

CHAPTER 1

Introduction

1.1 Motivation

When I speak to people from outside the scientific community of the field about brain-computer interfacing (BCI), one of the first questions I am usually asked is something like: "*So, can you guys already read my thoughts directly from my brain?*" The answer to this question of course depends to some extent on what the working definitions of "thoughts" and "directly" are in the particular case, but usually I can give the answer that this is not possible to date and it is also questionable whether it may ever be. However, the fact that this particular question is asked so frequently, gives expression to people's initial thoughts when they are confronted with the concept of BCI technology as well as an insight into how the potential scope, potential and impact of such a technology is perceived or imagined by the general population.

Moreover an increased public interest in the potentials and risks posed by BCI technology is reflected in its presence in popular television series and movies. For example, in the film *Elysium* from 2013, the character Max uses a BCI to access and control a highly advanced exoskeleton, allowing him to perform superhuman feats of strength and endurance. The BCI allows him to seamlessly control the exoskeleton and carry out tasks that would be impossible without it. In the film *Surrogates* from 2009, people use BCI technology to control and merge with robotic bodies - called surrogates - that allow people to experience their lives without the risks and limitations of the physical world. The surrogates enable people to live out their dreams and desires in a virtual environment, and they also provide a means for people to interact with others in a safe and controlled manner. Another example is portrayed in the episode *The entire history of you* of the popular series *black mirror*, which in general covers futuristic scenarios related to advancements in different technologies. Here a sensor device is implanted into the back of people's heads. Through this device a person's memories are recorded and their experiences can

be replayed on a “redo mode”. In this episode the purpose of the BCI is said to reduce crime, terrorism and simplify human relationships through transparency. In a different episode, called *Playtest*, the protagonist agrees to participate in the test phase of a video game, which blurs the line between reality and fiction. To play the game a chip is implanted into the player’s neck and linked to his neural system to adapt the game play to his most horrific phobias. This example further highlights a growing interest of merging virtual reality (VR) with BCI technology. The potential social and ethical impact of BCI technology is covered further in the very recent series *Made for love* from HBO released in 2021. Here the female protagonist wants to escape her unhappy marriage, but realizes that her tech billionaire husband has implanted a new tracking device into her brain - allowing him to access her current thoughts as well as her memories.

Futuristic scenarios like these give further insight and conclusions about the (perceived) possible potentials and risks emerging through advances in BCI technology. Today however, we are still a far way from direct encoding of a person’s thoughts and emotions or neurological hacking by collecting and analysing their patterns in assessed brain activity. Rather, today’s scientific investigations into BCI application are largely focused on medical assistance for impaired populations, as well as improvement of human-machine interaction through adaptation to a user that is based on more general mental state monitoring through neurophysiological measurement and its interpretations through a BCI. Additionally, most experiments investigating BCI applications are taking place under well-controlled conditions in a laboratory, which makes it questionable whether achieved results in e.g. classification accuracy can be replicated in complex and uncontrolled real-world scenarios. Above that, there also exist a number of other challenges which need to be addressed for feasible application of BCIs in future real-world applications and scenarios - regardless of whether the technology may be frightening and utilised for control and manipulation, or to increase safety, fun and performance in the interaction between humans and machines. Three of these challenges - *subject-dependence*, *task-dependence* and *increasing versatility through novel useful neural correlates for passive BCI applications* - I examined in three distinct studies in this thesis, investigating potential methods and paths to overcoming them.

1.2 Scope and structure of this thesis

The approach of *Neuroadaptive Technology* (NAT, see section 2.6.3) utilizes *passive* BCI (see section 2.6 and 2.6.3) for detection of specific aspects of the current cognitive user state (see section 2.6.2) and relates them to the given context. That way, the human brain, naturally interpreting the given context, can potentially deliver data

containing information about the context useful for statistical inference through machine learning. This can enable a machine to learn over time how the human user perceives, analyses, and interprets the world around them, and the machine's own actions can be adapted accordingly. This neuroadaptive approach holds great potential for the future development of intelligent machines and their application in human-machine interaction (HMI) contexts. Yet, today substantial hindrances and challenges for the application of passive BCI technology in real-world contexts remain and must be accounted for in order to enable future neuroadaptive applications.

In this work I firstly detail the state of the art of (passive) BCI and the current challenges standing in the way of real world applicability for BCI technology (Chapter 2). Following this overview, I present three distinct studies in the course of this thesis which address three current challenges for real-world application of passive BCIs.

I firstly address the problem of *subject-dependence* (Chapter 3). Today, when the predictive model of a BCI is trained, training data must be collected beforehand from the specific person, who later wants to use the trained model in an (online) application with new incoming data. This step of subject-specific training data collection is often time consuming and impractical for most HMI contexts. I therefore investigated the training of a *subject-independent* predictive model, which is trained on a group of other people's data, which then is applied to online-testing data of a completely new person. This investigation of a subject-independent classifier is performed in the context of training a classifier for (automation) error detection and correction in a real-world context.

In a second study, which was already published in Andreessen et al. (2021), I investigate the problem of *task-dependence* (Chapter 4). In order to gather training data for calibrating the predictive model of a passive BCI, usually in what is known as a calibration paradigm, the user completes a number of trials with activities that should elicit the cognitive signal of interest in them, as well as other trials where the signal is not generated. Further, for some mental states, e.g. mental workload or relaxation, different levels of the respective mental state, e.g. high, medium or low, can be induced through the trials of the calibration paradigm. On the basis of this calibration data, the classifier model is then trained to distinguish between data in which the neural activity of the cognitive signal of interest is present and in which it is absent, or to differentiate between different levels of the mental state in a more continuous way. Such a calibration paradigm is usually generated for every specific

context, task and application where a passive BCI is trained for online application - hence developing calibration paradigms is a time consuming and not very generalized process. However, some tasks that users are performing potentially are based on similar cognitive processes, making it unnecessarily effortful to design a specific calibration paradigm for each of them. Based on promising findings from Krol et al. (2016), I tested a potentially task-independent mental workload calibration paradigm. While this proposed paradigm is based on calibration data from the subject performing an arithmetic task, I applied the resulting classifier model in a new speed reading context. To this end, subject-specific predictive models were trained after subjects completed this task-independent mental workload calibration paradigm. The resulting predictive model for general mental workload was then applied to monitor the subject's mental workload while they performed the specific task of reading text material of different difficulty and reading-speed.

In a third study I examine the challenge of increasing *versatility of passive brain-computer interfaces through utilization of new neurophysiological correlates of mental states* (Chapter 5). In this study I investigate whether a predictive model of a passive BCI can be trained to distinguish between subjectively morally unacceptable and morally neutral stimuli, on the basis of the measured neurophysiological activity evoked by according stimuli. A passive BCI capable of classifying and monitoring human moral assessments could e.g. be employed in moral user modelling for future human-machine (moral) value alignment. Moreover, it would add human moral assessment to the possible user states which can be measured and monitored by a passive BCI. This would broaden the scope of contexts and applications in which passive BCIs could be employed.

Finally, I will discuss the findings from these three studies and their limitations in the context of stepping toward real-world application of passive BCIs and NAT. Additionally, an outlook on future topics to consider and challenges to overcome will be given in order to enable future applications of passive BCIs for improved interaction between human and machine (Chapter 6).

CHAPTER 2

State of the art: Brain-Computer Interfaces (BCI)

2.1 Preliminary note

In this chapter I will propose an overview of the state of the art and important concepts of (passive) BCIs and I will detail the methodological approaches which are used throughout the presented work of this thesis.

Firstly, I will introduce the general concept of BCIs and the components which belong to it. Subsequently I present methods for measuring neurophysiological signals with a main focus on electroencephalography which is widely utilized in BCI research. Afterwards I give an overview of the distinct signal processing and machine learning steps in BCIs and present how the concept of *passive* BCIs enhances human-machine interaction and enables neuroadaptive technology. Further I will describe current challenges for passive brain-computer interfacing in real-world human-machine interaction contexts.

The chapter concludes with an overview of three of those challenges which are addressed in the three distinct studies conducted within the framework of this thesis, namely (1) *subject-dependence*, (2) *task-dependence* and (3) *increasing versatility of passive BCIs by finding new neurophysiological signals* which can be utilized and assessed by a passive BCI.

2.2 Definition and components of BCIs

A *brain-computer interface* (BCI) can be described as a communication system - a system which establishes a direct communication and/or control channel between the brain and the environment (Wolpaw and Wolpaw, 2012). This communication system is solely based on brain activity without any muscular intervention between a person and an electric or mechanical system (e.g. a software application, a robot,

exoskeletons, voice prostheses or a wheelchair). Jacques Vidal was one of the first to use information from brain activity to detect a user's *mental state*. His work determined that brain signals are related to a person's movements and intentions (Vidal, 1973).

Historically, the idea of BCI systems originated in developing a means of communication for *clinical* contexts (Wolpaw et al., 2002). Probably the most significant application of BCIs in clinical contexts has been developed for groups of patients with *amyotrophic lateral sclerosis* (ALS) who have no or limited ability to control their muscles. This illness can be so pronounced that affected individuals are incapable of any voluntary muscle activity and all forms of communication are impossible – as in so-called *locked-in* patients. Here BCIs were intended to establish a novel communication channel that would enable patients to interact with the environment or control assistive technologies through signals generated via the brain (Kübler et al., 2005). In addition, BCIs can be used to control neuroprostheses and thus replace bodily functions (Müller-Putz et al., 2005) or in neurorehabilitation, e.g., in stroke patients, to restore cognitive functions (Shih et al., 2012; Buch et al., 2008).

Probably the most prominent example of such communication through a BCI is the *P300-speller* (Farwell and Donchin, 1988), which enables patients to spell out letters and words through interpretation of their brain activity. In this BCI, a matrix of all letters of the alphabet is presented to the patient where rows and columns are sequentially illuminated. Patients are asked to view the matrix and observe whether the desired target letter illuminates. From the continuously measured brain activity data, the letter is subsequently identified by eliciting a *P300* potential (Farwell and Donchin, 1988; Sellers and Donchin, 2006; Krusienski et al., 2008) - an *event-related potential* (see section 2.4.1). However, If patients are exposed to visual stimuli for an extended amount of time, they may become fatigued and have aching eyes (K.-T. Kim et al., 2016), and some patients lose voluntary control of the eyes in the late stage of the disease (Hinterberger et al., 2004; Sharma et al., 2011), which makes the speller an insufficient (long term) aid to some patients (Pasqualotto et al., 2015).

Although the original purpose of using BCIs was to help disabled or impaired patients, to date BCI application outside of the clinical context for the general population is often the focus of development, as commercial interest in such a technique has developed. In addition to clinical applications, today e.g. the gaming (Nijholt et al., 2009; Bonnet et al., 2013), education (Spüler et al., 2017), cognitive enhancement (Kaimara et al., 2020), safety management (J. Chen and Song, 2016; Nakanishi et al., 2013), (mental) health (Lim et al., 2012; Zotev et al., 2014), automotive (Lin et al., 2010; Khaliliardali et al., 2012), aviation (Aricò et al., 2016), marketing (Vecchiato et al., 2011) and defense industries (Kotchetkov et al., 2010) are engaging in BCI

research. It should be noted that the greatest potential for healthy users to date is associated with the subcategory of *passive BCIs* (see section 2.6.1), which enable user state monitoring for detection of e.g. attention, workload, fatigue, stress as well as error perception, without requiring attention or other mental resources from the user. However, BCI technology in general is still in the developmental stage, facing challenges (see section 2.7), such as low level of accuracy, efficiency, usability, as well as speed of BCIs (Van Erp et al., 2012). Moreover, BCIs need to be not only tested in the laboratory, but rather in real-world environments (Minguillon et al., 2017; Vařeka and Ladouce, 2021).

In general, the implementation of a BCI is divided into two parts: one part that concerns the measurement of the brain activity of a person – the *signal acquisition* (see section 2.3) - and another part which requires an analysis and handling of the obtained neurophysiological signals – the *signal processing* (see section 2.5). The signal processing usually consists of a pre-processing, feature extraction, classification and model application phase. Figure 2.7 illustrates these different parts of a BCI. In the following sections, both, signal acquisition and signal processing in a BCI, will be presented.

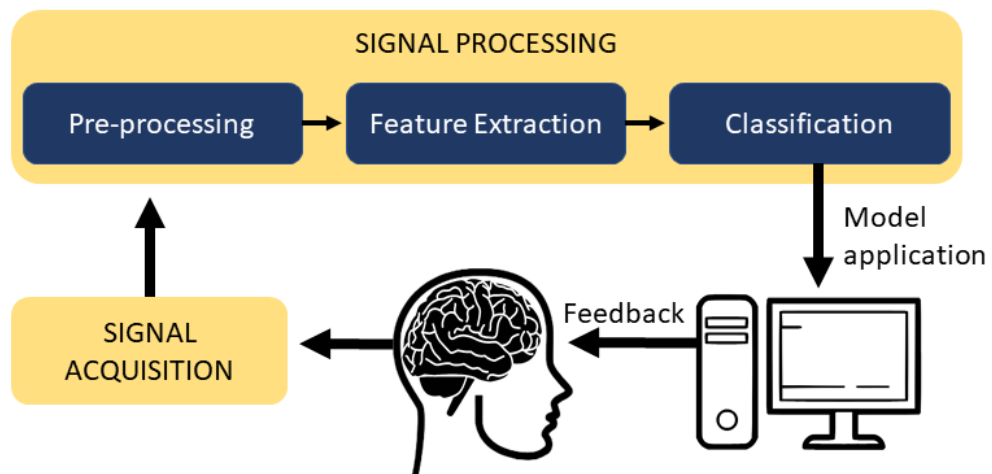


Fig. 2.1 Typical components of a BCI: Neurophysiological activity is recorded from the user's brain during signal acquisition. During signal processing the acquired data is pre-processed, features are extracted and classification is performed. The classifier model's output is finally used to generate commands to the system to provide feedback to the user or for control or adaptivity of the system.

2.3 Signal acquisition

Neurophysiological data can be obtained from brain signals in an *invasive* or a *non-invasive* way.

As for the invasive (or direct) method, the neuronal signal on the axon of an individual neuron or the electric activity of a group of neurons is measured directly from the cortex. This technique involves surgical intervention to place the measuring sensors under the scalp, as e.g. for *electrocorticography* (ECoG) an electrode array is positioned either epidurally or subdurally (Thongpang et al., 2011). Invasive techniques deliver a very precise signal characterized by a very good spatial and temporal resolution (Ramadan and Vasilakos, 2017), as they are picking up changes in electric activity directly from the signal's origin of source within the brain. Currently the high spatial resolution of this technology, however, is often of very localized nature, as implanted electrodes or arrays only pick up information from the areas over which they have been placed without recording much information from other areas of the brain. Invasive techniques are mainly applied in the clinical context, as their advantages are counterbalanced by far-reaching ethical challenges and safety-legal aspects (Rao et al., 2012). In addition, neurosurgical interventions are generally associated with health risks as well as high costs (Shih et al., 2012).

Non-invasive neurophysiological measurements are taken from outside the skull and do not require surgery for putting the sensors into place. As those sensors (e.g. electrodes) which are recording changes in neurophysiological activity are further away from the neural source eliciting the signal, it is not possible to collect the activity of a single neuron, but only of synchronously active groups of neurons. This also means that the *signal to noise ratio* (SNR) is lower than in invasive measurement methods. Moreover, in long-term measurement, non-invasive measurement techniques may have stability problems in sensor placement on the scalp. For (passive) BCI applications for the general, healthy population, non-invasive measurement techniques are of main interest, as they do not require surgical intervention and therefore present little risk to the user or patient. For an overview of non-invasive measurement techniques please see Castermans et al. (2013).

In the most common non-invasive measurement technique, namely the *electroencephalogram* (EEG – see section 2.4) electrodes are placed on the scalp for measuring brain activity in form of changes in emitted *electric* signals. However, it is also possible to monitor the *hemodynamic* activity of the brain by means of functional magnetic resonance imaging (fMRI) or near-infrared spectroscopy (NIRS) and magnetic encephalography (MEG) that are based on the analysis of magnetic fields (Gürkök and Nijholt, 2012). With regard to everyday real-world use for mea-

surement of brain activity, neither MEG nor fMRI are suitable methods due to their size and weight of the hardware (Arico et al., 2016).

Due to its high temporal resolution, affordable cost, and portable nature, the predominant method for measuring brain activity in BCI research is the EEG (Rak et al., 2012) which is described in the next section. For an overview of temporal and spatial resolution of different neurophysiological measurement techniques see figure 2.7.

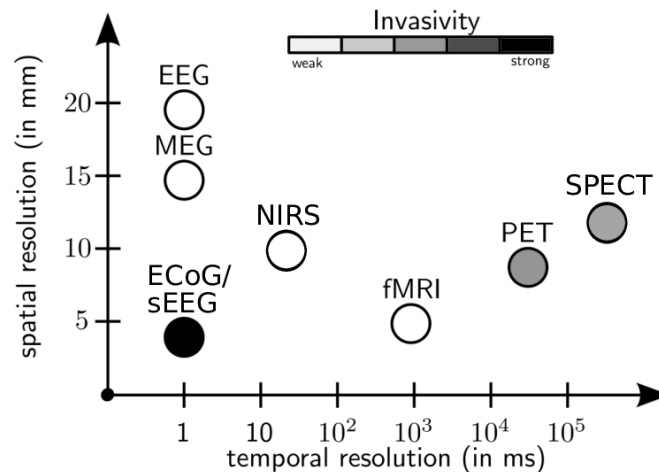


Fig. 2.2 Overview and comparison of different invasive and non-invasive neurophysiological measurement techniques by spatial and temporal resolution. *Figure source: Hitziger (2015)*

2.4 Electroencephalography (EEG)

Electroencephalography, first described by Berger almost 100 years ago, offers the possibility of recording the electrical activity of the brain by noninvasive measurement of electric potential fluctuations at the scalp in real-time (Berger, 1929). The signals are generated by potential fluctuations in the cerebral cortex where the nerve cells of the brain produce so-called action potentials through their electrical state changes for information processing. Here, only the simultaneous activity of many neurons of a certain area is measurable through the EEG method, because the potential changes that individual cells emit are too small, and only the summation of these potential changes can be measured from the outside the skull on the scalp (Zschocke and Hansen, 2011). Typically, oscillations with frequencies between 0.1 and 100 Hz and amplitudes of about 1 to 100 μ V can be measured in the EEG. Today, EEG is the standard method of examination in neurology and in clinical investigations.

In the BCI field, especially for applications intended for the healthy population, the use of noninvasive methods dominates, with EEG being the most commonly

used measurement method. The EEG method has a higher temporal resolution than many other measurement methods (e.g. fMRI) in the range of milliseconds (Teplan et al., 2002), is more portable than devices for most other measurement techniques and is comparably low in cost (Thomas and Vinod, 2017). Another advantage of the EEG method are extensive findings and insights from basic neurophysiological science from almost 100 years. Over the years several studies have demonstrated that an individual's mental state is correlated with certain characteristics of their EEG signals, both in terms of the timing and frequency of the signals (see next two sections 2.4.1 and 2.4.2). However, spatial resolution of the EEG signal is relatively low, with a high susceptibility to *artefacts* (Lakshmi et al., 2014). Artefacts are potential fluctuations that are not triggered by the user's brain. Both technically induced interference signals (e.g., line noise) and person-related factors (e.g., eye blinks or muscle movements) are sources for artefacts in the obtained signal recorded by EEG (Tandle et al., 2015).

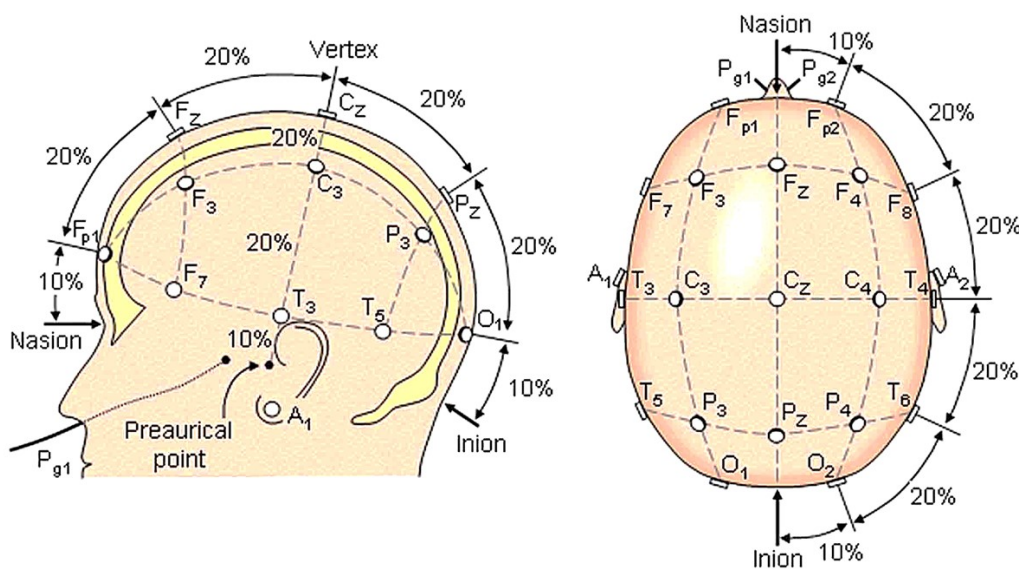


Fig. 2.3 The international 10/20 system for EEG electrode placement. Electrodes are named according to their sites on the scalp with F for frontal, P for parietal, A for anterior, T for temporal, C for central and O for occipital sites. Electrodes towards the left side of the head are labeled with odd numbers, electrodes to the right side with even numbers and electrodes on the midline are labeled with a Z at the end. *Figure source: Shriram et al. (2013)*

Regarding the setup of an EEG system, the *sensors* - the electrodes of the EEG - are usually mounted on a fitted EEG cap on the person's head and electrolyte gel is injected into the electrodes to improve skin conductivity on the scalp. Electrode placement is determined by a standardized system, the so called 10/20 system (Jasper, 1958) (see figure 2.3). This system determines the specific locations

for the placement of electrodes on the scalp by drawing lines between the nasion (the top of the nose) and the inion (the base of the occipital bone at the back of the head) and between the two preauricular points (the bones located in front of the tragus, a small cartilage fold in the external ear). A grid is then generated on the scalp based on these lines, with electrodes placed at the points of intersection within grid. The distance between each electrode is either 10 percent or 20 percent of the overall length of the line (Palumbo et al., 2010). The original 10/20 system contains 21 electrodes and other extended versions of the system include up to 345 positions for electrode placement (Oostenveld and Praamstra, 2001).

The electrodes of an EEG system are generally connected to an amplifier which enhances the recorded signal, and is again connected to a machine or computer for data recording and processing. Variations in the capabilities and features of available system for recording EEG signals are typically related to the effectiveness of the amplifier, the precision of the digitization process, the quantity of electrodes included, and the amount of data points that the system can measure and record per second (sampling rate in *Hz*).

In contrast to wet electrode EEG systems which require application of conductive gel, in recent years so-called dry electrode EEG headsets have been developed. The advantage of *dry* electrode systems is a significantly lower preparation time as gel application is omitted. Instead of gel application, often pressure is used to establish the connection between the electrode and the scalp (Zander et al., 2017). However, in many dry EEG systems the signal quality is much lower compared to recording with wet electrodes (Tautan et al., 2014) and prolonged pressure to the scalp can lead to discomfort in users, such as headaches (Mathewson et al., 2017; Zander et al., 2017).

In the analysis of the EEG signal, spontaneous spectral changes in neural activity and event-related potentials (ERPs) are distinguished. Both are described in the next two sections.

2.4.1 Event-related Potentials (ERP)

Event-related potentials (ERP) are small potential fluctuations in the EEG which occur at defined intervals in response to an external stimulus or event (sensory, cognitive, or motor) and result from its processing in the brain (Luck, 2012). The potentials derived at the brain surface represent electrical voltage differences and are displayed in their temporal course as curves in the microvolt range. The two primary factors that are considered when analyzing ERPs are the *amplitude*, which is measured in microvolts (μV), and the *latency*, or the amount of time between the

start of the stimulus and the maximum amplitude measured in milliseconds (ms). Both factors enable assessments of the strength and progression of the underlying processes and functions that occur within the brain.

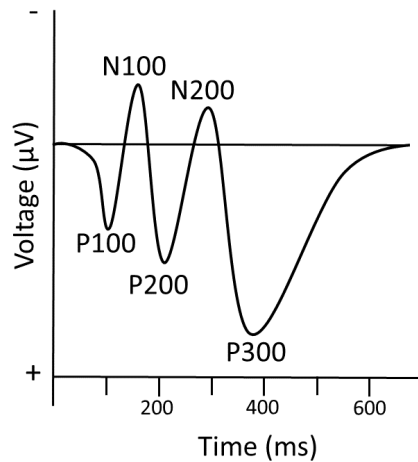


Fig. 2.4 Resulting ERP wave form with ERP components after data epochs following each presented stimulus are extracted and averaged together to create the averaged ERP, including the P300 component among other components. *Illustration according to Luck et al. (2000)*

ERPs occur through particular activation of cerebral sources and do not have a unitary amplitude; therefore, EEG recordings of ERPs can be superimposed by *spontaneous* EEG activity that is unrelated to the neurophysiological reaction to the event of interest (e.g. activity from other processes in the brain). Hence, in order to calculate and visualize the stimulus-dependent amplitude changes relating to the cognitive process of interest, the recorded EEG curves containing the ERP are subjected to an *averaging technique* (averaging). These averaging procedures are based on the assumption that the spontaneous EEG is subject to fluctuations into both the positive and negative amplitude ranges with respect to an event while ERPs are fairly constant with each repetition of the same stimulus. To obtain an average curve of all runs, the potential fluctuations following an event are measured, added, and the sum divided by the number of all events of the same type - the *trials* (Luck, 2014; Picton et al., 2000). In this way, the random background noise of the EEG can be filtered out. With an increasing number of trials, the ERP becomes more discernible, since a larger number of trials increases the probability that the values from the spontaneous activity balance each other out. The different amplitude excursions of the ERP, i.e. their maxima and minima, are referred to as the ERP's *components*. These components are named according to the polarity of the amplitude: "P" for positive and "N" for negative. In addition, the latency with which the component occurs is also specified - e.g. N200 (see figure 2.4).

A prominent example of an ERP used in BCI research is the so-called *P300* component (Donchin and Cohen, 1969; Pritchard, 1981). This component has a positive polarity and occurs with a maximum over fronto-central sites with a delay of about 300 ms after stimulus presentation (see figure 2.4), for example, when a series of the same, task-irrelevant auditory tones is followed by a different, task-relevant tone (this is the so-called *oddball paradigm* (Segalowitz and Barnes, 1993) which is also utilized in the P300-speller mentioned in section 2.2). The P300 component as an ERP is mostly used in *reactive* BCIs (for categorisation of different BCIs see section 2.6.1). An ERP often used in a *passive* BCI (see section 2.6.1), the BCI category which this thesis is focusing on, is the *error potential* (see section 2.6.2), which occurs when a person perceives an error (Chavarriaga and Millán, 2010). Such an error can be either committed by the user themselves or by another entity, such as another person, machine or by an application the user is interacting with. In recent years many BCI studies were performed where a classifier was trained to detect this error potential in the EEG data, and the classifiers output was then e.g. used to enhance performance of the user or the machine or to adapt the human-machine interaction (Parra et al., 2003; Zander et al., 2010).

2.4.2 Spectral changes in neural activity

In the analysis of spectral changes in neural activity, the oscillations (see figure 2.5) in the measured EEG signal are divided according to frequency into five types of waves - the *frequency bands* - which are typically associated with specific mental states and brain functions (Constant and Sabourdin, 2012) as shown in table 2.1.

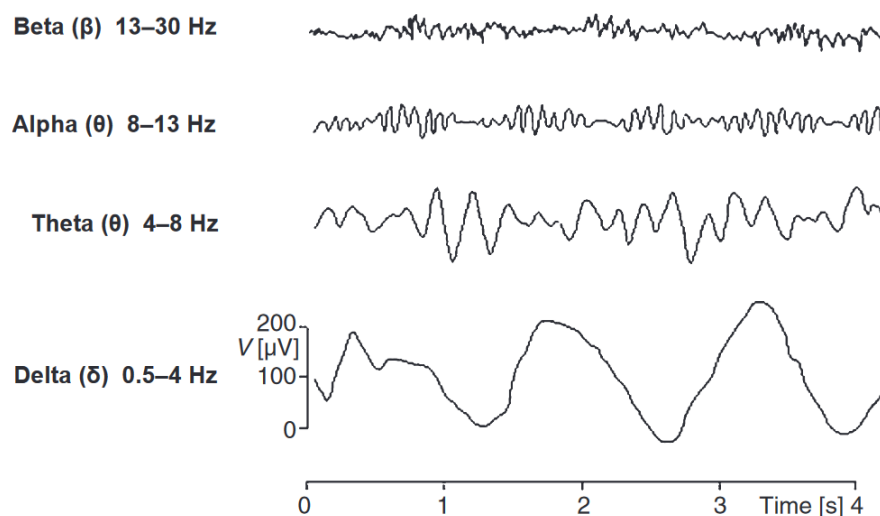


Fig. 2.5 Exemplary oscillations of the frequency bands in the EEG. *Figure source: Constant and Sabourdin (2012)*

Gamma waves have the highest frequency (>30 Hz) and are measurable during learning and attention processes of the brain. *Beta* waves are characterized by rapid oscillations of 13–30 Hz and are associated with states of wakefulness and alertness in a person who is actively thinking. In the relaxed or meditating awake state with closed eyes, the EEG synchronizes at a frequency of 8-13 Hz, the so called *alpha* waves, which are particularly pronounced above the occipital pole. With the onset of the sleep state, waves with greater amplitude and lower frequency occur: *theta* waves (4-8 Hz) during drowsiness and light sleep phases and *delta* waves (0.5-4 Hz) in dreamless deep sleep phases or comatose states.

Table 2.1 Overview of the five frequency bands in the EEG and examples of corresponding mental states in which they occur according to Constant and Sabourdin (2012)

Frequency band	Frequency range (Hz)	Corresponding mental state
Delta	0.5-4	dreamless deep sleep phases, unconscious or comatose state
Theta	4-8	drowsiness and light sleep phases
Alpha	8-13	awake state with closed eyes, relaxed, or meditating awake
Beta	13-30	normal waking consciousness, alertness, aware of self and surroundings
Gamma	>30	most active conscious waking state with higher mental activity as e.g. learning and attention processes

In BCI research an often applied approach using spectral changes in EEG activity is the *event-related desynchronization* (ERD, Pfurtscheller and Da Silva (1999)) of the mu-rhythm in *motor imagery* paradigms (Pfurtscheller et al., 2006). The *mu-rhythm* lies within the alpha frequency band at 8 to 12 Hz and is highest over the sensorimotor cortex, a brain area which is associated with a person's coordinating motor action and perception. In an idle state brain activity in this area is synchronized and therefore a high power in the mu-rhythm can be measured. When motor actions are imagined by the person (e.g. repeatedly forming a fist with the right or left hand) the mu-rhythm over contralateral corresponding sites of the sensorimotor cortex decreases due to increased cortical activation. This means that imagining movements of the right hand would result in an ERD over the left sensorimotor cortex which corresponds to movement of hands. ERD-based BCI paradigms where movements of distinct body parts are imagined, are usually applied for navigational

control over a device or a software application as e.g. navigation of a wheelchair (Tsui et al., 2011) or controlling movement of a cursor (McFarland et al., 1997).

While ERD-based BCIs belong to the category of *active BCIs* (see section 2.6) as the user generates the mental state the BCI operates on themselves - by actively imagining limb movements - a much investigated approach of a *passive* BCI where spectral changes in neural activity are employed are BCIs that detect different levels of mental workload. It is generally observed that as mental workload increases, there is a corresponding increase in activity in the theta frequency band over frontal electrodes, while activity in the alpha frequency band simultaneously decreases over parietal sites (Gevins et al., 1997; Krause et al., 2010). BCIs measuring mental workload are further discussed in section 2.6.2 and the second study of this thesis in chapter 4.

2.4.3 Independent Component Analysis (ICA)

When EEG is used for recording electric activity from the scalp, the signal every sensor - each electrode - is recording from their specific location is the sum of a mixture of sources which emit electric activity. These sources can be cortical or non-cortical, hence not only brain activity is contributing to the measured EEG signal, but also electric activity from other sources within the body, such as eye movements and blinks, heart rate or other muscular activity. Further electric signals from the external environment can also contribute to the measured signal, such as electric devices or line noise. As mentioned before in section 2.4 these two categories of non-cortical activity are called artefacts in the EEG, as they do not represent the cortical activity of interest. Overall, the signal measured at a specific time point at a sensor is the sum of electric activity from all those sources combined - activity from cortical sources as well activity from artefactual sources (Jung et al., 2001).

Independent component analysis (ICA) is an efficient method for signal separation, which can be applied to investigate the EEG signal's underlying original sources (of activity), their activity time series, and in which way these sources separately contribute to the recorded activity on the scalp (Makeig et al., 1995; Jung et al., 2001). An ICA model (see figure 2.6) is able to reconstruct the sources of a mixture of signals, as ICA calculates statistically independent linear combined components, under the assumption that the data are maximally statistically independent if the data are non-gaussian distributed.

The EEG data measured in μV can be described as a matrix X , with a row for each channel (electrode) and a column for each sampling point - this is the *channel space*. The ICA is then used to estimate a filter matrix W which enables us to unmix

the measured data from the channel space (X) into a resulting matrix of *independent components* (ICs) C – the *source space*.

$$C = WX$$

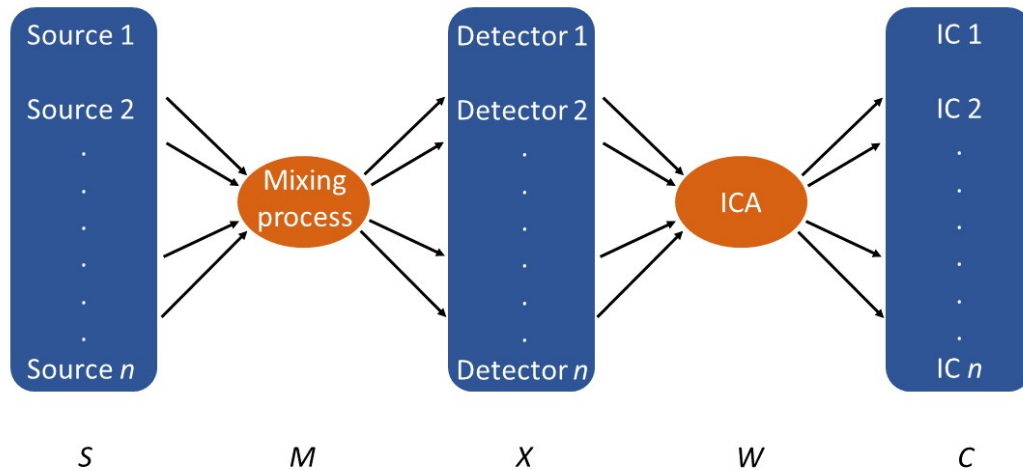


Fig. 2.6 Overview of the ICA model. The n sources (S) of the signal are linearly mixed by an unknown process (M). The signal resulting from the mixed sources is recorded by a number of n detectors (X) - the electrodes of the EEG system (channel space). ICA is then used to estimate a filter matrix W which enables us to unmix the measured data from the channel space (X) into a resulting matrix of independent components (ICs) C – the source space. *Illustration according to Brown et al. (2001)*

By multiplying the filter matrix W with the observed data, the matrix of independent components (C) is obtained. The columns of C contain the activity profiles of the sources, which are maximally independent of each other.

$$X = W^{-1}C$$

The rows of W^{-1} analogously contain the weighting of the individual channels functioning as a *filter* (Brown et al., 2001; Jung et al., 2001), which allows e.g. to calculate the activity of a specific IC at a specific channel as well. These filters can be transformed into dipolar *patterns* representing the projection of single sources onto the scalp (Delorme et al., 2012). By using these projections, it is possible to determine the location of sources within the three-dimensional space that is covered by the channels (Delorme et al., 2012). Moreover, the filters in combination with back projection also allow for removing activity from artefactual ICs from the overall measured sum of activity in the channel space. If an unwanted or artefactual

IC (e.g. heart activity or eye movements) is identified, the *activation* of this IC can be extracted and subtracted from the overall signal by setting the corresponding rows and columns of the matrices to zero. When back-projected into the channel space the activity of the artefactual IC is then removed from the overall measured activation in X . This way, ICA is increasingly being used as a suitable method for artefact reduction in EEG for further analysis, as well as a method for identifying and studying temporal and spatial features of cortical sources (single and clusters of ICs) underlying neurophysiological phenomena in the neuroscientific field and brain-computer interfacing.

2.5 Signal processing and machine learning steps in a BCI

As outlined in section 2.2 the functioning of BCIs consists mainly of the two steps of signal acquisition and signal processing. After introducing concepts and methods for signal acquisition in section 2.3 where neurophysiological activity is recorded from sensors and then amplified, I will now detail the different steps of signal processing in the following sections. These essential steps include the following four phases in a BCI:

- **pre-processing:** this step aims at reducing noise resulting from artefacts and improving the quality of the signal
- **feature extraction:** at this step, the signal's relevant characteristics are identified and extracted. These characteristics of the signal - the features - are pertinent to the task of classifying the brain state, and they contain the signal variations that are related to the user's underlying neurophysiological processes. These features will be used later to control or adapt the system through the BCI by interpreting them in the context of the user's brain state
- **classification:** the feature vector extracted during feature extraction is given as an input to the classifier approach. The classifier then learns in a predictive model how to associate the vector to different classes of data through machine learning approaches
- **model application:** finally mental states are classified and identified through the predictive model of the classifier. The classifier is applied to new incoming data and its output is then either translated into a command (to e.g. control a machine, or generate a feedback signal) or processed otherwise (e.g. for adaptive adjustment of a system)

Finally, I will summarize how BCI approaches can be validated after a predictive model has been trained for a BCI application.

2.5.1 Pre-Processing

The purpose of signal pre-processing is to improve the *signal-to-noise ratio* (SNR) and to highlight relevant information contained in the recorded signal.

After acquisition the raw recorded data contains noise and artefacts which are not of cerebral origin. Hence the recorded data is a mixture of the brain signal which contains the information relevant for the classification (information relevant for the mental state or task) and the noise which represents all other superimposed signal components, including physiological signals, such as heart rate, muscular activity (e.g. from the jaw) and eye movements or blinks. In general the amplitudes of these sources of noise are much higher than those of activity stemming from the brain. Moreover, technical artefacts can contribute to the recorded EEG, such as line noise (at 50 or 60 Hz depending on the location) or noise induced by electrode movements occurring during the data recording.

During pre-processing the afore mentioned artefacts are removed from the EEG data as well as possible. For increasing the SNR, *filters* can be applied to the data. These filters can be spatial, reducing the number of electrodes from the original number to a subset of electrodes that are expected to contain relevant information about the signal of interest. As an example, in the context of a BCI that uses ERD for detecting motor imagery, the electrodes that are most relevant are typically C3 and C4, that are positioned over the left and right motor cortex. These electrodes are used to identify changes in brain activity related to the user's imagined movements. Also *common average reference* (CAR) can be applied as a spatial filter to reduce overall background noise. The CAR method includes taking the average of the recorded signal of all channels and then subtracting it from each channel's signal. This process results in the inclusion of only signal and noise that is common to all channels in the CAR, while signal that is isolated to a single site is not included (Osselton, 1965; Ludwig et al., 2009). Hence, uncorrelated random noise is minimized through the averaging process by the subtraction, leaving channels with only channel specific signals (Teplan et al., 2002).

Frequency filters can also be set - mostly *low-pass* (all frequencies below a selected threshold are preserved), *high-pass* (all frequencies above a selected threshold are preserved) or *band-pass* (frequencies between two thresholds are preserved) filters, as well as *notch*-filters (all frequencies but those between two thresholds are preserved (is often used to remove power line noise at 50 or 60 Hz)) are set to limit data processing to the frequency ranges where the neurophysiological signals are

presumptively present. In general, most of the signals that are of interest in EEG data fall within the frequency range of 0 Hz to 40 Hz. The above filters can also remove slow drifts from the EEG signal, as well as power line noise.

Apart from more simple spatial or frequency filters, also ICA, as described in 2.4.3, can be used to clean data from artefacts by removing data generated by artefactual components from the EEG data. In case other physiological data was measured during the recording, such as eye activity through an *electrooculogram* (EOG), this data can also be used to separate data derived from neuronal sources from data caused by artefacts. During pre-processing other methods can be applied to the data as well such as *down-sampling* for reduction of data dimensionality, or a *baseline correction* can be performed, to correct for potential *non-stationarity* in the EEG signals - the change of characteristics of a time series over time (such as variance, mean, and spectral properties) which are apparent between the initial measurement of the EEG signals for the purpose of calibrating the BCI and the actual use of the BCI in an experiment or application, due to changes in the user's attention, fatigue, task, substance intake or stress etc. (Blanco et al., 1995; Shenoy et al., 2006). Further, data from faulty or noisy EEG channels can be removed from the EEG data through *channel rejection*. Similarly, epochs contaminated by artefacts can be excluded through *epoch rejection* from the EEG data.

2.5.2 Feature extraction

The goal of *feature extraction* is to extract distinct information about the relevant event or mental state which is to be detected from the EEG data through the BCI system. To this end the aim of this phase is to build filters for the data which describe specific aspects of this activity, the so-called *features*, which can be used for classification.

After the EEG data has been partially cleaned from artefacts and noise of non-cerebral signal sources during the pre-processing phase, in feature extraction characteristics of the cerebral signal of interest are extracted and separated from other cerebral activity. The recorded brain signals, which are represented as a matrix with spatial (electrodes) and temporal (time stream of recorded data points in ms) dimensions, are mapped onto vectors containing the representative features which best represent those aspects of brain activity relevant to the corresponding neurophysiological event or user state. This new data space can be very different from and more abstract than the original spatial-temporal structure of the raw EEG and is referred to as the *feature space*. Class related relevant features that are to be represented in the feature vectors are usually identified from neuroscientific literature. In alignment to the research question and the type of data at hand, features

can be selected from the temporal or spectral domain, or a combination of both. Data from outside the feature space are excluded from further analysis. It should be noted that high dimensionality of EEG data can be reduced drastically by feature selection, which can have a positive effect on later classification and BCI performance robustness.

There are several methods for feature extraction such as e.g. common spatial patterns (CSP) or auto-regressive (AR) - for an overview see e.g. Pawar and Dhage (2020). In the studies within this thesis, the following two approaches for feature extraction were used:

- In studies 1 and 3 ERP-based BCI systems are trained, which are based on features from the *temporal domain*. For feature extraction we used a time-point based approach based on averaging over voltages of time-points within a specific time frame of the data following stimulus presentation. To this end *epochs* starting around 100ms before stimulus onset and with a specific length are extracted from the data. Within these epochs time-windows (e.g. of 50ms length each) are defined in which characteristic properties of the signal of interest are expected (e.g. for a P300 ERP component). Then all data points within a time-window are averaged for each electrode. The feature vector for each specific electrode then is defined by the time-windows and their average voltage. In the BCILAB toolbox used for data analysis in this thesis this time-point based feature extraction procedure is implemented in the *windowed-means approach* (Blankertz et al., 2011) for data processing, feature extraction and classification.
- In study 2 the BCI system which is trained, is based on spectral features of frequency band power. Here the *filter bank Common spatial patterns* (fbCSP, Ang et al. (2008)) approach was used for feature extraction. fbCSP is an advancement of the CSP approach for feature extraction, where spectral, as well as spatial discriminative EEG characteristics are selected and optimized. Firstly, EEG signals are filtered into a number of frequency bands by applying a filter bank. Then, for each selected frequency band, CSP features are extracted by applying spatial filters in order to maximize the variance of one class while minimizing the variance of the other class. Afterwards discriminative pairs of frequency bands and corresponding CSP features are chosen by the algorithm of the approach.

2.5.3 Classification

After pre-processing and feature extraction, the goal of classification for a BCI is to train a *classifier model* through machine learning techniques (Bishop and Nasrabadi, 2006) to distinguish between EEG data which is elicited by the event and/or mental state of interest and other EEG data. To this end a machine learning algorithm is applied to learn how to separate data from two (or more) classes by building patterns containing class specific features. Ideally, the classifier model after training is able to distinguish between these classes with a low error rate on a *single-trial* basis in real-time. The latter point is essential for BCI technology operating on EEG data, as the afore mentioned averaging over trials (as e.g. for neuroscientific ERP analysis) is not possible with a BCI. Rather, the challenge is that the event- or mental state-relevant information must be obtained directly from the recordings of a single trial. The information is to be extracted as patterns from the EEG data, based on pre-denoted predictive models of the BCI classifier. The trained classifier later should be able to perform pattern matching and decide for an unlabeled and new data point, which class it belongs to according to its features.

During its *calibration* the classifier algorithm usually is provided with labeled data containing information on what part of the data belongs to which class. In order to obtain these single-trial examples of the signals to be detected in the mixed scalp EEG for the calibration of the classifier's predictive model, labeled examples of brain data elicited by the target stimulus or mental state within the user are collected beforehand. In most cases, this calibration data is acquired while the user completes a *calibration paradigm*. This calibration paradigm usually consists of a number of different trials which represent the data *classes* for later calibration. Part of these trials – in which the ground-truth is assumed - are designed to evoke the brain activity of interest (e.g. elicit an ERP (as e.g. the P300 or an error potential) or induce the mental state of interest (high mental workload, high alertness). The other class(es) of trials a user completes during the calibration paradigm either aim to induce a contrasting mental state - e.g. if there are two types of calibration trials, of which one evokes high levels of mental workload while the other evokes low levels of workload or relaxation while the aim is to train a BCI classification model which can distinguish between low and high stages of mental workload. Other classes of trials can also simply serve as a *baseline* where no specific mental state is induced or distractor stimuli are presented to the user. A typical setup of the EEG hardware, subject and calibration paradigm in a laboratory experiment is shown in figure 2.7. Preconditions for good classification results are a sufficient number of samples and reliable data. The number and/or length of trials a user has to complete for each class depends on the number of independent features ex-

tracted, the number of classes the classifier later will be trained to distinguish, as well as the machine learning method applied for classifier training, as some methods require more samples for robust classification than others. It should also be noted that the tasks presented to the user in the calibration paradigm should be as similar and comparable to the task or mental state the BCI later should be able to detect - this should always be considered when designing a calibration paradigm for collecting training data for a BCI, especially when it should be applied in real-world settings later on.

To calibrate a BCI, patterns must be extracted and learned from training data that allow the properties of the data to be categorized in order to make predictions based on them. The goal is to recognize patterns in the data and to derive regularities from them, which can then be applied generally to the categorization and processing of new data. Hence, after sufficient training data of the different classes has been collected, pre-processed and features were extracted, in the next step the actual classification of the data follows to obtain a *predictive classifier model*. The algorithms used for the classification are selected on the basis of the data structure and the problem. The goal is to perform an assignment of single trial data to classes based on the feature vectors, which can be transferred into binary or more commands. Pattern recognition or classifications are performed on the previously described features.

Among other machine learning algorithms for classification (as e.g. neural networks or support vector machines) *linear discriminant analysis* (LDA, (Fisher, 1936)) is commonly used for classification in BCI as it is robust and of low computational cost and therefore applicable for online application. Hence, LDA was employed as a classifier in the experiments detailed in this thesis, and it will be discussed in more detail in the paragraph that follows. For an overview of other classification algorithms for BCI technology and their characteristics see Lotte (2008), Lotte et al. (2018) or Müller et al. (2003).

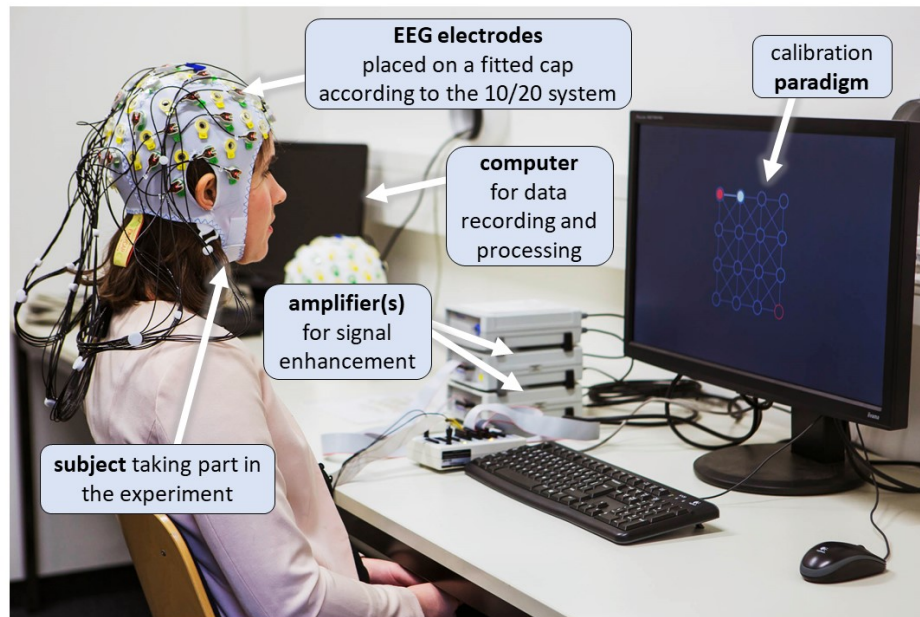


Fig. 2.7 Typical setup of a BCI experiment in the laboratory: The subject is seated during the experiment. On a screen the calibration paradigm is presented to the subject, where exemplary data for the classes of brain activity of interest is evoked and collected. EEG electrodes are mounted to the head of the subject with electrode placement according to the extended international 10/20 system on a fitted cap. The electrodes are connected to one or more amplifier(s) to enhance the EEG signal. The EEG signal is then recorded and processed on a connected computer. *Picture source: TU Berlin/PR/Oana Popa-Costea*

Linear Discriminant Analysis (LDA)

With the use of the classification method LDA, it is possible to identify a linear combination of attributes that distinguish or characterize two or more classes of objects or events (Bishop and Nasrabadi, 2006). A linear classifier can be created and applied using the resultant combination.

The optimization procedure of LDA is based on covariance estimates of feature distributions where the LDA uses a linear hyperplane to assign data to a class (Duda et al., 2012). This linear hyperplane attempts to maximize the distance between the means of the classes and minimize the variance within the classes (Duda et al., 1973). The results are reflected in a linear transformation vector, which operates as a filter that isolates the relevant activity and reduces unrelated activity. The class of a new sample or trial is then accordingly determined by the side of the hyperplane it is placed on. As mentioned above, LDA as a linear classifier is low in computational cost and therefore applicable for online application. Moreover, linear classifiers are more robust and are less prone to *overfitting* (where the classifier model becomes overly focused on unique or erroneous differences between the calibration

data classes) than non-linear ones (Vapnik, 1999a). It should be noted though that LDA performance can be impacted when the number of training samples is limited while the size of the feature vector is high (Blankertz et al., 2011) - this problem can however be counter measured by applying shrinkage for compensating the resulting bias of the estimated covariance matrices (Blankertz et al., 2011). This approach is then called (shrinkage) regularized LDA.

After its calibration a classifier can be tested *offline* to obtain an estimated accuracy for online BCI performance when new, unlabeled data points are to be assigned to a class. A common method to estimate this accuracy of a trained classifier model is *cross-validation*, which is described in the next paragraph.

Cross-validation

Cross validation is a common statistical method often used in BCI research to estimate the goodness of online performance of a trained classifier, meaning how well it classifies data on which it has not been trained. Therefore the calibration data is divided into several small data sets. In this process, a partial dataset is used for testing and the remaining partial datasets are used for training. For example, a 5×5 cross-validation divides the data into five partial data sets. One partial data set is used as the test set, and the other four are used as the training set. In order to utilize each individual subset as a test set once, this is done five times. The process is repeated five times, so that for the objective evaluation of the classifier 25 test sets and 25 training sets are available. In the end classification accuracies obtained from testing on all testing datasets are averaged to define an overall performance measure which estimates the classification accuracy of the classifier for online application. A threshold for acceptable accuracy of a specific classifier can be defined by a method from Müller-Putz et al. (2008) in which it is determined whether a classifier model operates *better than random*.

In the context of a BCI the classifier is trained on calibration data and then applied to the data produced during an application. Overall, classification by the classifier's predictive model, the filters, can be used *offline* to learn about the user's relation to a context, task or environment or *online* to drive control or adaptivity of an application in real-time, as is described in the next section.

2.5.4 Model application and validation of BCI approaches

After a classifier model has been trained and its goodness was estimated offline, in its application phase the classifier model is applied to new data samples in order to assign these to one of the classes. The new data samples are usually generated

while the user is interacting with a system in real-time, and the aim is to interpret and detect changes in the user's state or the user's perception of events from classification run on the ongoing online EEG data stream. The classification *output* is then finally interpreted and can be translated into commands, feedback to the user or for adaptation of the device or application, mainly in real-time.

The generated filters of a classifier model trained with LDA do not automatically differentiate between brain and non-brain activity. As a result, the filters are the best for class classification, but they cannot be read spatially from a neuroscientific viewpoint. Systematic eye blinks, for example, may be included in the filters if ocular activity contributes to class separation. As a result, if we want to identify true and robust brain correlates of specific user states using BCI, we must take care to ensure and verify that the BCI operates on the basis of relevant brain activity - of the cortex or other brain regions. The classifier's *spatial filters* can be transformed into *spatial patterns*, which can provide insights into the scalp projections of cortical activity that is relevant for the classification process (Haufe et al., 2014). For each feature, such a pattern that is interpretable from a neuroscientific perspective can be created, identifying the activity at specific electrodes that contributed to distinguishing between classes within a specific frequency range. This process allows for the identification and validation of the information utilized by the classifier in the discrimination of classes, as analyzing the spatial patterns of brain activity can aid in determining the underlying sources that contribute to the functioning of a BCI. For instance, the impact of eye movements can be easily identified based on their unique spatial patterns.

While spatial patterns can be interpreted in sensor space to validate if a BCI operates on cortical activity, the activity displayed in the spatial patterns is that of a mixture of the activity of underlying sources – the individual independent components whose activity is contributing to the classifier's decisions. For further validation of a BCI, it can however also be of interest to investigate the contribution of these individual sources. As described in an approach for a classifier source localisation tool by Krol, Mousavi, et al. (2018) and Zander et al. (2016), a combination of ICA sources – the independent components resulting from ICA - and a spatio-temporal LDA-based classifier model can be employed to identify sources whose activation is used to discriminate between classes by the classifier model. Through this approach a classifier model that was trained in sensor space can later be interpreted and analyzed in source space. To that end, the weights of the LDA's spatial filters can be transformed into weights in source space, which represent the ICs' relative contribution to the process of classification (Zander et al., 2016). Through these new weights in source space, it is possible to visualize the relevance of cortical areas to the classifier model by generating *weighted dipole density plots* as also de-

scribed by Krol, Mousavi, et al. (2018). Through this approach the classifier model of a BCI can be validated further and the underlying neurophysiological processes can be compared to already existing findings from neuroscientific research.

2.6 Passive BCIs

In this section I will firstly introduce the concept of *passive* brain-computer interfaces (passive BCIs) and how they form a separate category of BCIs that differs from *active* or *reactive* BCIs. I will further give an overview of the fields of application for passive BCI technology and present EEG features and correlates of user states that are commonly assessed and applied in passive BCIs for user monitoring in HMI contexts. Finally, I will introduce the concept of Neuroadaptive Technology (NAT) which as a technology is enabled by passive BCIs.

2.6.1 Categorisation into active, reactive and passive BCIs

BCIs in general can be categorized according to different characteristics (for example, synchronization or degree of invasiveness). One categorization is concerned with what a user needs do in order to utilize the BCI. To that end, BCIs are divided into three categories based on the type of the signals used to generate input commands from the user's point of (Zander and Kothe, 2011; Zander and Jatzev, 2011):

- **Active BCIs:** the BCI's output is derived directly from a conscious (control) thought of the user, independent of external inputs. With these BCIs control is *intentional* from the user perspective, as the analyzed signal is consciously modulated by the user.

An example for an active BCI is the motor imagery paradigm, where the user intentionally imagines limb movements to generate control commands to e.g. control a wheelchair or a cursor. The imagination of a movement is accompanied by activation in the region of the motor cortex that is also activated when movements are actually performed, and can be interpreted by the BCI accordingly.

- **Reactive BCIs:** here the BCI's output is derived from the user's internal reaction in brain activity to an external stimulus or to an event. This type of BCI requires the user's directed attention to the stimulus or event presented to them, as its output is generated by actively modulated perceptual processes within the user. Hence, as with active BCIs, a reactive BCI requires the user's conscious decision to send a command.

An example for a reactive BCI is the P300 speller discussed earlier, in which rows and columns of letters are flashed in quick succession with the user's focus set on the letter they want to spell out.

- **Passive BCIs:** the BCI derives its output from the user's naturally occurring, spontaneous changes in brain activity within a specific context, without the user having to deliberately communicate anything. As a result, the user remains passive in terms of consciously providing input to the BCI system because their brain activity is not intentionally generated or intended to operate as an input signal or command to the BCI system. A passive BCI, however, uses these measurements of spontaneous brain activity to automatically infer knowledge about the user's mental state; this type of input is known as *implicit* input (Zander et al., 2014). The interpreted information about the user's mental state can then be applied to adapt the interaction between the user and a device or an application. As the user neither has to actively or reactively control one specific application or machine through the BCI, many passive BCIs can run in parallel to assess different aspects of a user's state.

One example for a passive BCI could be the ongoing assessment of a student's mental workload levels, that are derived from spectral changes in measurements of their brain activity, while the student is learning tasks for a math test. In order to make the learning process more effective, the exercises could be simplified automatically if the mental workload is too high (overload) according to the passive BCI's output, and the difficulty of the exercises could be increased if the mental workload is too low.

In this thesis the focus lies on passive BCIs in particular, as passive BCIs currently hold the greatest potential for enhancing and improving *human-machine interaction* (HMI) applications for the healthy population on a broad scale. The primary aim of passive BCIs is to detect changes in a user's state and use this information as an input for adaptive systems (Aricò et al., 2016) – as will be discussed in the next section.

2.6.2 Passive BCIs for improved Human-Machine Interaction (HMI)

The main goals of BCI research over the past two decades have been to define support systems for people with severe physical impairments, e.g. stroke or locked-in patients, that allow control over a technical system or can be used as a communication channel (Birbaumer et al., 1999; Kübler et al., 2005). The expansion of categorization of BCI technologies to include the concept of passive BCI (Zander and Kothe, 2011) has opened up a new field of application for BCIs in the field of

HMI. A new channel of interaction with technology has been established, in which brain signals may be recorded and analyzed within a specific context while the user interacts with a system, and their brain activity is interpreted so that the system's behavior can be adapted to the user. As detailed in section 2.6 passive BCIs do not aim to directly control the system through intentional commands from the user, but rather to extract information about "covert aspects of [the] ongoing user state" (p.2, Zander and Jatzev (2011)) that may be provided to the machine to improve the underlying HMI. Such an implicit input channel could be usefully added to ongoing HMI in the future without increasing the user's cognitive effort, since the use of a passive BCI inherently does not require the user's attention.

The above described form of human-technology communication aims to develop context-sensitive technical systems that proactively support the user. George and Lécuyer (2010) describe such *implicit interaction* as "an interaction process that is not based on direct, explicit, or voluntary action of the user, but more on the state of the user in a particular context. Both the user's state and the given context can thus be associated with the expression *implicit information*" (p.2). Implicit interaction can be achieved by allowing system behavior to be adapted to the user's needs and expectations in real-time, informing and improving the ongoing human-machine interaction while taking the user's context into account (Zander and Jatzev, 2011). Thus, a passive BCI can be regarded as a means to interact implicitly with a machine based on neurophysiological activity. Additional assessment of eye tracking e.g. can provide further contextual information about visual input and events perceived by the user which can be combined with information about the user state which is derived from passive BCI output (Protzak et al., 2013). Hence, passive BCI technology has the potential to address diverse application markets.

In the following I will present EEG features reflecting neural correlates of cognitive user states that are commonly assessed and applied in passive BCIs to provide implicit information about the user. I will further refer to exemplary studies in which these EEG features were investigated and a classifier model implemented for potential non-medical, as well as real-world passive BCI applications and case scenarios. For an extensive overview of these potential areas for application of passive BCI technology please refer to e.g. Blankertz et al. (2010) or Alimardani and Hiraki (2020).

- **Error detection:** An *error-related potential* (ErrP) is an ERP measured in the EEG that is related to the perception of errors (Falkenstein et al., 2000). This ERP generally shows as an early negative and a later positive potential - the *error-related negativity* (ERN) (Gehring et al., 1993) which is peaking at round

50-200 ms after stimulus perception over fronto-central sites and the *error-related positivity* (Pe) (Falkenstein et al., 1991) which shows a centro-parietal maximum around 200 to 500 ms after an error was perceived (Kumar et al., 2019). Neuroscientific studies found that both components, the ERN and the Pe, are most likely generated in the *anterior cingulate cortex* (ACC) (Dehaene et al., 1994; Miltner et al., 1997; Taylor et al., 2007). The ACC is an area of the brain which has been found to be generally sensitive to conflict appearing among the representations generated while a task is performed (see e.g. Carter et al. (1998)).

It should be noted that the perception of an error can arise from a variety of scenarios, all of which elicit an ErrP with small neurophysiological variations as a response (Ferrez and Millán, 2005; S. K. Kim et al., 2017). In the context of HMI, perceived errors can e.g. be grouped as follows (see S. K. Kim et al. (2017) or Spüler and Niethammer (2015)):

Automation error: the output of a system or an application can differ from the user's expectation (*automation error*). Here the user observes an error committed by a third party - namely the system or application they are interacting with - hence this is also called the *observation ErrP*. Moreover, the user may encounter a situation where the machine is not responding to a command in the way intended by the user - this is called the *execution or interaction ErrP*.

User error: a user can commit (unintentional) errors while interacting with a system (*user error*). Here the user can realize themselves that they committed an error - this would be the *response ErrP* - or the user is informed by some kind of feedback that they committed an error - this would be the *feedback ErrP*.

Both - automation and user errors - can lead to user frustration and negatively impact the interaction between a system and a user (Zander et al., 2010). For a further overview of differences in the variety of errors and how factors as the type of error modulate the production of the ERN and Pe components of the ErrP please refer to Spüler and Niethammer (2015), Scheffers and Coles (2000), Wirth et al. (2019) or Falkenstein et al. (2000).

The ErrP has been studied in multiple EEG and BCI studies for a variety of scenarios and allows to identify, predict or even prevent errors in real-time within HMI contexts (Zander et al., 2010; Chavarriaga et al., 2014). Such an error-detection passive BCI allows for the development of adaptive systems

which increase performance and avoid frustration in the user. For instance, a *hybrid* BCI approach (Pfurtscheller et al., 2010) can be created by integrating the ErrP signal with traditional BCIs, where a reactive BCI, in form of a P300 speller, is combined with a passive BCI for error-detection. The ErrPs detected by the passive BCI can be employed for corrective action to improve the speed and general efficacy of communication in the BCI speller (Schmidt et al., 2012; Spüler et al., 2012).

In an early BCI study on ErrPs Parra et al. (2003) had subjects perform a visual discrimination task. When subjects committed an error, they were given a feedback by the system. An error detection classifier was trained for each subject resulting in high single-trial accuracies with an average of 89 percent. Online application of the trained passive BCI error-detection models to correct for errors committed in the manually given responses by the subjects yielded a significant reduction in errors for most participants of around 21 percent.

Other studies also monitored ErrPs which were then used as a corrective during HMI interaction. Kreilinger et al. (2012) e.g. attempted to monitor the accuracy of the movements of a robot controlled arm by assessing error potentials.

Yousefi et al. (2018) had subjects perform different kinds of tasks (mental arithmetic, counting, word generation, figure rotation and idle state). Errors in the interaction with the system were introduced by erroneous feedback given after completion of some trials. The resulting ERP from erroneous trials showed an ErrP with a negativity at ~ 180 ms, and subsequently two positive deflections at ~ 400 and ~ 630 ms after the feedback onset. Classification between correct and erroneous feedback trials showed an average of 83 percent accuracy.

ErrPs can also be used to adapt to a user's intentions, as an adaptive system provided with feedback through an error-detection passive BCI can learn from its mistakes over time, which may result in a reduction of misclassifying the user's intentions. In a corresponding human-robot co-adaptation scenario Ehrlich and Cheng (2018) used ErrPs as a feedback signal that can be indicative for human preferences. Here, a co-adaptive agreement between the human and the machine was necessary for an interaction to be deemed successful and the robot altered its actions based on ErrPs detected from the human user's ongoing online EEG. Average error-detection of the trained passive BCI classifier model yielded 82 percent and the study resulted in successful human-robot co-adaptation that created a more efficient interaction be-

tween human and robot.

Additionally, De Visser et al. (2018) investigated whether ErrPs measured by the EEG could be used to establish an indicator of *trust* in automation that does not require self-report by the user. They recorded EEG from 21 participants who watched different algorithms carry out a flanker task. Each algorithm had different characteristics of credibility and reliability. The results showed that ErrPs were elicited as participants monitored the observed algorithms and that the Pe can be applied to differentiate between levels of reliability in the algorithms and that it significantly correlates with subjective measures of trust.

The focus within this work, and in the later presented first study in this dissertation (see chapter 3) lies on automation error detection, specifically the detection of executions errors in a realistic HMI scenario, as I investigated session- and subject-independence for classification in a passive BCI for error-detection.

- **Mental workload assessment:** As already described in 2.4.2, many studies have investigated the correlation between levels of mental workload and measures of theta band power over frontal electrodes and alpha band power over parietal electrodes in the EEG (Gevins et al., 1997; Krause et al., 2010). While frontal theta power was typically shown to be positively correlated with mental workload, the correlation between parietal alpha power and mental workload is negative (Holm et al., 2009). Recent studies have achieved good classification accuracies for passive BCIs that are trained to assess low and high mental workload states of a user and predict such levels of mental workload in different task domains such as e.g. visual-spatial task (e.g. mental rotation; Berka et al. (2007), span tasks (Chaouachi et al., 2011), arithmetic tasks (Walter et al., 2017), n-back tasks (Grimes et al., 2008; Herff et al., 2014) and Sternberg tasks (Baldwin and Penaranda, 2012; Roy et al., 2016).

For HMI contexts a continuous assessment of mental workload through a passive BCI can play an important role for a variety of contexts as e.g. safety-critical environments or learning environments. Here the system a user or operator is interacting with, could be adapted according to the passive BCI's assessment of the (current) workload level by e.g. decreasing task difficulty or complexity when high levels of mental workload are detected.

In research regarding mental workload assessment in HMI contexts, specifically in aviation, e.g. Arico et al. (2016) tested a passive BCI that assessed ongoing mental workload in order to adapt the interaction between a pilot

and the automated cockpit for increased flight safety and reduction of errors caused by human operators.

In a fNIRS study, Gateau et al. (2018) used a passive BCI to monitor mental workload under real flight conditions in aircraft pilots to discriminate two levels of mental workload during aircraft piloting tasks. They yielded relatively high accuracies of more than 76 percent in mental workload classification, demonstrating the feasibility of monitoring mental workload through a passive BCI in a realistic and complex situation. In a similar setup Dehais et al. (2019) monitored two levels of mental workload under real operational flight conditions using a 6-electrode dry EEG headset. Results showed higher alpha band power over parietal areas and higher theta band power over the occipital area in the low mental load condition compared to the high load condition. Classification accuracies for these frequency features yielded an average of 70 percent.

Schultze-Kraft et al. (2016) assessed mental workload in human operators in a real-life workplace environment, yielding average classification accuracies of around 94 percent.

Kohlmorgen et al. (2007) presented mental workload assessment through a passive BCI in a real-world driving scenario, where the user was engaging and completing tasks with a car's electronic warning and information system while operating under real traffic conditions. If levels of high mental workload were detected in the user through the passive BCI, the current task was interrupted to mitigate the induced mental workload in the user. Results showed that the users' overall task performance increased through the adaptive system.

In an adaptive learning scenario, Yuksel et al. (2016) employed a passive BCI which was assessing mental workload through fNIRS measurement to adjust a musical training task's level of difficulty. Here, the system automatically increased the task difficulty when workload levels of the monitored pianist user fell under a pre-defined threshold, indicating that the practiced material had been learned and enough cognitive resources were available in the pianist learner to be introduced to a new part of the practiced piano piece. Results showed that learners who were supported by the passive BCI played with significantly increased accuracy and speed, and subjectively felt that they learned better through the support of the adaptive passive BCI system.

In this thesis, mental-workload assessment through a passive BCI is performed in the second study (see chapter 4), where the task-independence of a mental

workload classifier is investigated in the context of a text reading task.

- **Attention and task engagement:** Changes in attention within a user or human operator can impact the safety and efficiency within HMI operations. The measurement of drops in attention through a passive BCI has been investigated in different studies e.g. in the context of aviation or industrial workspaces where fluctuating concentration levels in human operators who are conducting passive, monotonous control tasks can create dangerous situations. In such situations passive BCI based monitoring of attention could be used to decrease task related strain on users e.g. by decreasing task speed, mitigating distractions or demanding that the user takes some time off the task to rest. Such a system which can adapt the interaction with the user in accordance to their attention levels would mitigate safety risks and allow for sustained user performance at optimal levels.

The afore described P300 ERP component (see 2.4.1) has been studied in a variety of BCI studies as a measure of attentional resources within a user. These studies showed that the amount of attentional resources available for processing an event corresponds proportionally to the amplitude of the P300 (Polich, 2007; Lew and Polich, 1993). Another marker for decreased attention are high activity in the theta band and low activity in the beta frequency band (Parasuraman and Rizzo, 2007).

Acı et al. (2019) used an EEG based passive BCI to monitor the three different mental states (focused, unfocused and drowsy) in five subjects. They found that frequency changes in the frequency ranges 1–5 Hz and 10–15 Hz at frontal and parietal areas were corresponding with changes in the subjects' state of attention and yielded average classification accuracies of 92 percent.

- **Relaxation:** A state of relaxation or meditation is usually reflected in increased activity in the alpha frequency band (Barry et al., 2007; Nowlis and Kamiya, 1970). Moreover, while in a meditative state of mind (which is characterized by internalized attention and mindfulness), practitioners of meditation show an increase of alpha and theta band power (Hebert et al., 2005).

Applications of passive BCIs which measure relaxation in a user have been implemented e.g. in some video games, in which specific features of the game are modulated by the measured level of relaxation in the user. For instance, Mühl et al. (2010) developed the multimodal *Bacteria Hunt* alpha neurofeedback game. In this game the player controls the movement of an amoeba with the keyboard. The aim is to move the amoeba towards a target - a fleeing bacterium - in order to eat it. EEG is used to monitor alpha power over parietal

regions in the player, as an indicator of relaxation. The measured alpha band power is then used to alter how far the bacteria targets can flee. Therefore, when a player is more relaxed, it becomes easier for him to reach the targets.

- **Automation surprise:** In situations where a conflict between the human operator of a machine (e.g. a pilot) and the automation system (e.g. the controls in an airplane cockpit) occurs, the operator can detect but not understand this conflict when they are surprised by the behavior of the automation. Such a system-output related situation, which causes confusion and distraction in the operator's mental state (e.g. through an unexpected specific alarm) is termed *automation surprise* (Sarter et al., 1997; De Boer and Dekker, 2017). Automation surprise can impact safety through the elicited confusion and attentional shift in the operator. Therefore states of automation surprise and their causes should be detected in order to avoid critical situations and improve human-machine interaction (Roy et al., 2020; Dehais et al., 2015). Automation surprise can be detected in the EEG as an ErrP, even though it is not an error in the automation per se, but rather results from a human-machine coordination breakdown, resulting from a deviation between expectation of the operator and actual system behavior (De Boer and Dekker, 2017). As described above, an ErrP generally consists of a negative ERP component at fronto-central electrodes, followed by a positive deflection at centro-parietal areas.

Pawlitzki et al. (2021) e.g. investigated the calibration of a passive BCI for the detection of automation surprise, as well as errors in an aviation context. Through a calibration paradigm, in thirteen pilots cognitive states which corresponded to surprise and/or error were evoked. Three according classifier models were then trained per subject, one for classification of automation surprise, one for error detection and one for surprise detection. The three classifier models were afterwards applied to EEG data recorded during a simulated flight task which included events that corresponded to surprise, error and automation surprise (as assessed explicitly by the individual pilot). The average classification accuracy for automation surprise in this study was 85 percent. Regarding the classifier models' specificity, the BCI models for automation surprise and surprise detection resulted in similar output, while the error detection BCI model produced output significantly different from the other two classifiers. Pawlitzki et al. (2021) conclude, that both surprise as well as error contribute to the mental state of automation surprise.

- **Detection of emotions:** The detection of different states of emotion is another feature in the EEG which has been researched in recent years. This re-

search is closely connected to the field of *affective computing* (Picard, 2000), which aims at allowing machines to recognize human emotion, react to them and ultimately express emotions themselves. Affective computing however includes other physiological measures aside from brain activity as well, such as heart response, facial recognition or speech. However, passive BCIs for emotion detection can be applied as an input channel for the detection of emotion, in affective computing and other domains.

In a review of studies aiming at identifying different emotional states in humans, Al-Nafjan et al. (2017) found that in most of such investigations "elicitation techniques include standardized emotional stimuli (e.g. pictures, films and audio), imagination techniques (e.g. guided imagery and autobiographic recall), present social interactions (e.g. games) and directed facial action tasks" (p.7) to gather classifier training data in a calibration paradigm. Regarding the neural correlates measured by EEG for emotion detection, in the same review Al-Nafjan et al. (2017) found that out of the 285 publications included in the review analysis, "130 articles used event-related potentials, whereas 48 articles used frontal EEG asymmetry in their analysis, six articles used ERD/ERS and four articles used *steady-state visually evoked potentials*" (p.10) (SSVEP, see Zhu et al. (2010)).

In order to distinguish six different emotions, namely happiness, surprise, anger, fear, disgust and sadness, Nasehi et al. (2012) developed an algorithm which extracts spectral, spatial, and temporal characteristics from EEG recordings related to emotions. The trained classifier correctly classified emotions at an average accuracy of 65 percent. It should be noted that here the chance level is different from binary classification (50 percent), as six different emotions were classified in a senary classification, hence the chance level would be at 16.7 percent.

Chanel et al. (2011) investigated emotion assessment from physiological signals to adapt the difficulty of a game - namely Tetris. The three emotional states which were assessed in this study were boredom, engagement and anxiety. Classification results from only the EEG signal yielded an average of 56 percent accuracy and could be raised by including data from other peripheral physiological features to 63 percent.

Anh et al. (2012) used a classifier to distinguish between five emotions elicited by presenting pictures from the *International Affective Picture System* (IAPS, Lang et al. (1997)) to participants. The induced emotions, based on valence and arousal scaling, were happy, angry, sad, relaxed and neutral. Classification of these emotions in real-time and achieved accuracies of an average of 70

percent.

Daly et al. (2016) created an affective brain-computer music interface that might change its user's emotional states during interaction. The final online application of the classifier could detect current emotional states within a user with classification accuracies of up to 65 percent.

Though there is a variety of studies which investigated automatic recognition of emotions through EEG and passive BCI systems, classification accuracies are usually low, making it questionable how viable emotion detection via passive BCI in real-world contexts can actually be.

As these different EEG features and their presented investigations for user state detection and monitoring show, there is immense potential for employment of passive BCI technology in a variety of fields.

The given examples already show that passive BCIs can potentially

- aid in improving safety and performance in operational workplaces, aviation and driving
- increase efficiency and personalization in learning setups
- increase user and/or system performance
- avoid errors and hazardous situations
- enable adaptation to a user's preferences
- avoid user frustration
- indicate trust in automation
- aid in improvement of another BCI's performance
- allow feature or difficulty modulation in gaming contexts and
- facilitate self reflection, increase mental health and training outcome in a user.

Apart from these factors for real-world application, analysis of the classifier models trained for passive BCI applications can also validate and enhance scientific insights from cognitive neuropsychology and related fields (Blankertz et al., 2016). Additionally, the application of BCIs for personal authentication is investigated (Chuang et al., 2013; Li et al., 2015) and neural markers are tested as indicators for psychological diagnoses for e.g. autism (Alsaggaf and Baaisharah, 2014) or ADHD

(Alchalabi et al., 2018) as well treatment for such conditions (Lim et al., 2012; Fan et al., 2015). Also artistic expression through a BCI, as in the perception and creation of art (Nijholt, 2019; Prpa and Pasquier, 2019; Krol, Andreessen, et al., 2018) and music (Miranda and Castet, 2014; Daly et al., 2016) through the generated output and interpretation of user states through a passive BCI is investigated and exhibited.

While passive BCIs can be regarded as a tool for user state assessment, the technological concept of HMI systems which are adapted through information about the user derived from neurophysiological measures in a closed-loop, while additionally capable of incorporating other context information, has been defined and researched under the term *Neuroadaptive Technology* (NAT). In the next section I will introduce this concept of NAT and how the tool of passive BCI is enabling this novel and promising technology.

2.6.3 Neuroadaptive Technology (NAT) enabled through passive BCIs

As introduced above, passive BCIs enable the detection of specific aspects of the transient cognitive user state and relate them to the given context. In this, the passive BCI is interpreting a user's mental state in an unobtrusive manner - the user does not need to allocate any mental resources to operating or monitoring the BCI system, but can fully focus on the task at hand or the experience they are currently in. By this means, the human brain, naturally interpreting the given context, can potentially deliver data containing information about the context useful for statistical inference (Vapnik, 2013) through machine learning. There trough, the machine could learn how the human user perceives, analyses, and interprets the world around it, and adapt its own actions accordingly. This approach leads to "a cybernetic convergence of human and machine intelligence" (p.5, Zander et al. (2016)) that is referred to as *Neuroadaptive Technology* (NAT). What these adaptations towards the user will look like, depends on the particular context of the application, as well as the task the user aims to perform. In particular, NAT works by identifying low-level data on brain activation (e.g. event-related potentials), which is then co-registered against the technology's repertoire of possible responses or states. The user's neurophysiological activity in response to known contextual changes is implicitly and continuously monitored, and this activity is cross-referenced in real-time with the changes that caused them. This enables the generative creation of a statistical model that represents the intentions and preferences of the user - the *user model*. The behavior of a system in a given circumstance may then be adjusted using this user model to take the user's current condition or particular goals into account. Moreover, the user model can be refined and updated over time.

2.7 Challenges for passive BCI applications in real-world contexts

Regardless of the potential benefits that passive BCI and NAT could bring to users in different industries and application fields, it is also necessary to consider current hurdles that need to be overcome before introducing and deploying these technologies. Even though over the past decade there was significant progress in research and development of potential real-world applications for passive BCIs for offline evaluation measures or as a modality for online adaptation in HMI contexts (see section 2.6.2), to date passive BCI systems are not mature enough to be brought to market or to be applied in real-world applications yet. There still exists a considerable number of challenges which will need to be solved before passive BCI can be established in everyday life or context specific scenarios to improve HMI, enhance human or machine performance or increase safety in critical environments.

In the following I will present some of these challenges for passive BCI applications. For an in-depth overview of challenges for future (passive) BCI application for real-world application, please refer to Rashid et al. (2020), Saha et al. (2019) or Ramadan and Vasilakos (2017).

In particular, I will introduce three specific challenges for passive BCI at the end of this section - subject-dependence, task-dependence and finding novel useful neural correlates for increased versatility of passive BCI applications - which are subject to the three studies I conducted during the course of this thesis in order to contribute to overcoming and finding solutions for these three obstacles in BCI research.

2.7.1 Sensor technology

In order to become applicable and acceptable to users in real-world scenarios and everyday life, the sensors (e.g. the EEG headset) used to collect neurophysiological data must become more wearable, as well as comfortable, affordable, should be wireless and probably more fashionable than the EEG systems available today, while not compromising signal quality of the acquired EEG data. More advancements should be made in the direction of developing gel-free (dry) electrode systems that do not require the application of conductive gel between each electrode and the scalp. This application of gel is too time consuming and leaves visible traces of the conductive gel in the user's hair and must be washed out - both these aspects of gel based EEG systems are preventing their everyday applicability and inhibit user acceptance. Furthermore, wired headsets limit movements, whereas wireless data-transmitting headsets would allow people to move without restriction.

In recent years several versions of newly developed commercial dry, as well as wireless EEG headsets were already developed and tested, e.g. regarding signal quality, wearing comfort, user acceptance and usability, as well as their robustness against sensor displacement and loss of scalp contact resulting from the user's body movements during task performance in real-world settings (Zander et al., 2011, 2017; Nijboer et al., 2015; Radüntz et al., 2019). The signal quality of dry EEG systems however remains inferior to gel-based systems and seems to be more prone to noise from e.g. movement artefacts or electrical interference (Saab et al., 2011), while in general a higher SNR in wet or dry sensor technology would be desirable. Moreover, dry EEG headsets often work with pressure towards the scalp to ensure maximal conductivity with the skin and a certain degree of stability in electrode placement, often resulting in discomfort for the user as e.g. headaches and other kinds of pain are induced after wearing the headset for some time. Additionally, even commercially available non-invasive EEG headsets often require some level of training to place the headset and the sensors correctly, and often require regular checking for contact quality to ensure signal quality and classification results.

2.7.2 Computation and classifier performance

For many real-world scenarios the classification accuracies for mental state detection reported in BCI literature is not sufficient yet (Évain et al., 2016). Hence, BCI technology in general is still much less reliable than other input devices. Advances in machine learning for improved classification and data handling are required to improve the performance of future BCI classification models, as user acceptance of the BCI technology may depend largely on users' perceptions of its utility and effectiveness (as well as comfort) (Évain et al., 2016).

Moreover, the resolution for measurement of a specific mental state should be increased to, e.g. enable measurement of workload at different levels. This would entail a more frequent shift from the prominent binary classification approaches in passive BCI towards multi-class classification (Venkatesan and Er, 2016).

The employment of high density EEG systems, to increase the spatial resolution of the measurement method, results in high computational cost (X. Chen et al., 2015). Here, to improve classifier accuracy and performance, it has been proposed to employ a combination of different neurophysiological measurement techniques, e.g. EEG and fNIRS, to exploit the advantages of high temporal and spatial resolution of the two different techniques in combination (Dehais et al., 2018). However, such a combination of techniques would also increase setup time, as two different systems must be set up. Additionally, wearing comfort for the user could be reduced through additional weight of the second system's sensor hardware.

Further, given how little data is often accessible prior to a classifier's application by a single user, methods must be created that can be calibrated or trained with the least amount of data (Fairclough and Lotte, 2020).

In order to identify and handle non-stationarities of neurophysiological data (see section 2.5.1) that decrease classifier performance over time, approaches like adaptive machine learning algorithms could prove useful to update parameters in the classifier model (Shenoy et al., 2006). Especially for online applications of passive BCIs and NAT algorithms must be created that are effective enough in terms of computing.

2.7.3 Artefacts and confounds

Most research on passive BCIs is currently conducted in well-controlled experimental settings, mostly in a laboratory with shielded rooms, where the user is shielded from environmental noise. Moreover, physical movement, such as walking or speaking, is often restricted during experiments in order to minimize the impact of movement related artefacts on the data recording. When passive BCI applications are taken out of the laboratory into real-world settings SNR can become lower as the amount of artefacts potentially included in the recorded data immensely increases, due to bodily movements of the user and other interfering electrical signals from the environment (Minguillon et al., 2017). Moreover, in real-life settings it is much more difficult to control for other signals generated through confounding mental states that occur simultaneously to the mental state that is to be assessed by the passive BCI. Parietal alpha activity e.g., which is an important feature in mental workload assessment, could become challenging to assess when a user is fulfilling tasks in a natural environment. The expected effect of a high workload environment on the measurement of alpha activity could be reversed or become undetectable due to changes in demand in the complex real-world scenario - e.g. when the focus of demand is regularly shifted between attention to a visual scene and memory retrieval, which also induces modulation in alpha activity (Blankertz et al., 2016). Thus, since many passive BCIs attempt to decode a particular mental state, it is necessary to find biomarkers that are unique to that mental state and do not vary in the presence and changes in other concurrent mental states in order to provide reliable decoding of the targeted mental state (Fairclough, 2009). It is however essential for a realistic investigation of the actual possibilities and challenges of passive BCIs in the real world, to study their application in scenarios with conditions as close to real life as possible. This facilitates in the identification of confounding factors that may impact neurophysiological variables in real-world contexts and may result in a decline in classifier performance outside of the laboratory. It will also aid in iden-

tifying characteristics of biomarkers that are *general* enough to be employed across different contexts, such as in the lab as well as in realistic settings with moving users in real-world environments.

Currently, procedures for online removal of artefacts, such as ICA, require a high amount of computational power, as well as time for data processing due to computational complexity and the high number of required channels - which induces time delays between the data recording and the resulting output generated by the passive BCI's classifier. For real-world scenarios where passive BCIs are to be employed in safety- and time-critical situations, such a delay is not feasible and unacceptable. Hence, future methods for artefact reduction in online applications of passive BCIs must be developed so they are able to function without or very little amounts of time required, as e.g. proposed in Mayeli et al. (2016) and Val-Calvo et al. (2019).

2.7.4 Legal, security and ethical considerations

Regarding ethical considerations, as well as data privacy and security issues, there exist concerns about potential "mind reading" abilities for future applications of BCI technology. Especially in form of passive BCIs which do not require the user's intention to communicate through the BCI but continuously monitor and interpret ongoing brain activity. As such information extracted from brain data may include information about "users' experiences, emotions, motivations, and other previously private mental and psychological processes" (p.1, (Schiliro et al., 2020)). Though BCI technology is not capable of direct mind reading yet, in some passive BCI studies it was found that the user was not aware that they had given the system any information (Zander et al., 2016) - this aspect may hold potential for manipulation and exploitation. Already in 1991, Farwell and Donchin demonstrated in the *Guilty Knowledge test* that a P300-based BCI can be used to identify whether a person recognizes items or images from e.g. a crime scene. In their experiment a person is asked to look at a series of images (such as from a crime scene). In case the person recognizes any of the images, a P300 response is evoked and can be read out from the recorded EEG data which in this case is an indication of *guilty knowledge*. While information like this would be desirable for investigative or law enforcement purposes, access to such brain data poses potentially serious risks to individual privacy rights.

Recently, advances in neuroscience, hardware, and software have led to the emergence of BCI systems in many other fields of application outside the medical domain. In the commercial sector, international companies such as Google or Facebook, which base their businesses on data collection, interpretation and trade, are taking steps towards assessing personal brain data from their users as well with

the prospect of even more precise and personalised target marketing. With international operations through these companies, questions arise regarding cross-border personal data transfers, compliance to ethical innovation and securing privacy rights regarding brain data. Questions as whether brain data can be sold, whether one needs to comply with certain rules and laws (such as the European GDPR) or what happens if a BCI system is hacked should ideally be addressed before functional applications of BCI technology are in place. Already to date, companies such as *Neurable* (<https://neurable.com/>), *Mendi* (<https://www.mendi.io/>), *InteraXon* (<https://chooseuse.com/>), *NeuroSky* (<https://www.brain-express.com/>), *BrainCapture* (<https://braincapture.dk/>) or *Dreem* (<https://dreem.com/>) operate their service by frequent collection of real-time data of brain activity from their users on their smartphones. For an overview of privacy and security risks posed by BCI technology please refer to Bonaci et al. (2014), Ienca et al. (2018) or McCullagh et al. (2014).

Data privacy and security issues that may arise from BCI technology and the collection of personal brain data in the future, should already be assessed today. To that end, scientists, companies and governments (at best on an international level) should confer the risks and consequences posed by BCIs and NAT to provide guidelines and laws regarding the handling and sharing of personal neurophysiological data. General frameworks for neurotechnologies and BCI as proposed by Chavarriaga (2020) or in a recent roadmap by IEEE (IEEE Standards Association, 2020), initiatives dedicated to assessing current requirements and difficulties for neurotechnology standardization, must be debated and updated to define standards for responsible innovation.

2.7.5 Subject-dependence

As mentioned before, to date, before a (passive) BCI can be applied within a specific (HMI) context to support a user or patient, due to subject-, session- and application-related differences (Shenoy et al., 2006) in the neurophysiological signal obtained by the EEG, a *calibration session* is required to collect user-specific training data for classifier calibration. In this calibration session, the user usually performs a number of known tasks inducing different stages of the mental states associated to these tasks. These collected and labeled examples of EEG activity, based on the user's specific, current brain activity, are then used to train the classifier predictive model to discriminate between the different tasks in accordance to the user's characteristic activation pattern (K.-R. Müller et al., 2004). However, this *subject-specific classifier model* needs to be retrained over time, normally before every new session where the subject-specific classifier model should be applied (online) by

the specific user. This recalibration becomes necessary due to the variability of the recorded EEG signal, which is e.g. caused by differences in sensor positioning between data recording- and model application sessions, as well as changes within the individual user between sessions and over time induced e.g. by mood, substance intake (caffeine, medication, drug use etc.), lack of sleep or stress (Yagi et al., 1999). Another related issue is the usage of different electrode setups and numbers that can be applied for the same passive BCI application - here a generalization across specific electrode montages without losses in classification accuracy would be desirable.

The need for a repeated calibration of the subject-specific classifier model before every use of the (passive) BCI system to achieve an optimal classification model for each user and each day poses a strong barrier for establishing practical and everyday usage of BCI technology. Calibration sessions usually require a few minutes of time or even longer. This, in many cases, is unacceptable in practice when a user wants or needs to be supported by the BCI on a daily basis, such as at work (it however always remains a question of cost and benefit, and for some applications repeated recalibration may still be acceptable). A BCI which could be used immediately and does not require recalibration for retraining the classification model, as well as data recording and system training for each new user, would be desirable for establishing everyday usage of BCI systems for the healthy population, as well as impaired patients.

In order to establish BCI usage on a larger scale and enable everyday NAT applications, the development of user-specific classifier-models which do not require retraining between sessions is crucial – such a classifier would be *session-independent*. Through machine-learning approaches, some session-independent active and reactive BCIs were developed e.g. for a P300-based BCI (Lu et al., 2009) or a motor imagery BCI (M. Krauledat et al., 2008), where data from multiple sessions of a specific subject was used for classifier training and the application of this session-independent classifier model to new online data did not result in significant losses in classification accuracy. But even if this problem of *session-dependence* can be resolved, every new potential user would still need to undergo at least one initial calibration session, before they could start using the BCI application for the first time. Here, it would be desirable to reduce the calibration time to zero for new users, by establishing a classifier model, which does not require the collection of calibration data from each new individual user - a *subject-independent* classifier model. Recently, unsupervised learning methods were proposed to construct such a subject-independent classifier model (Kindermans et al., 2014). In this approach, an underlying inherent structure is learned by the classifier by providing it with unlabeled data as input for classification – hence, no (labeled) calibration data recording is

needed. However, so far supervised techniques for classifier calibration are preferred since they still yield better classification accuracy than unsupervised learning approaches (Blankertz et al., 2016). Transfer learning (Jayaram et al., 2016) is a further strategy that was suggested for completely eliminating the requirement for new training data gathering for each new user. Here, shared structures between training data of multiple subjects and/or sessions are exploited to increase classifier performance in a new user. In such an approach towards subject-independent classification, the classifier model could be trained on calibration data previously collected from a larger number of other subjects and is then applied to the online data of the new user, whose data was not part of the calibration dataset the classifier model was trained on. For a motor imagery BCI e.g., Fazli et al. (2009) followed such an approach and trained a subject-independent classifier model for motor imagery on previously recorded data from 45 subjects. They found that through the use of the obtained subject-independent classifier model new users were able to use the BCI classifier in real-time with no further system calibration at low loss of performance levels. The rationale behind such an approach is that a classifier model based on training data from multiple subjects could provide better generalization properties in terms of the underlying neurophysiological processes related to the mental state of interest - compared to classifier filters based on just a single subject - and could therefore be used in new subjects without the need to recalibrate the system. This subject-independent classifier model would eliminate the calibration stage and thus enable plug-and-play application of BCI technology in many (HMI) contexts. Furthermore, a subject-independent classifier model, trained on data from a larger number of individuals, could provide insights into general human similarities in neurophysiological activity underlying mental states and processes as well - hence provide additional and novel knowledge of human neurophysiology.

In the first study of this dissertation (see chapter 3) I will explore the training of a session- and/or subject-independent passive BCI that can detect error-perception in a user. Specifically, I will follow the latter approach proposed above and aim to calibrate a subject-independent classifier model based on labeled data from a number of other subjects and then apply the obtained classifier model to data from a new subject to compare classification accuracy to that of a subject-specific classifier model for error-detection. At a different stage of the same study I will use subject-specific data from one recording session to train a subject-specific model and then apply this model to data from the same subject, but in a new session. Here I will examine if the trained classifier model can be applied session-independently in the new session and compare its classification performance to that of a subject-specific classifier model trained with data recorded in the same new session. An overview of the different parts (part 1 to 7) of this study, examining session- and

subject-independence in an error-detection passive BCI, is given in section 3.2.

2.7.6 Task-dependence

Patterns in assessed brain activity vary not only between individual users and recording sessions, but also between applications, tasks and contexts. Therefore, usually for every specific task or application where a BCI system should be employed, a specific training paradigm is especially designed to collect data which contains labeled, known examples of brain activity from the specific mental states the classifier model should learn to distinguish between. For example, when a passive BCI should be applied to assess mental workload while a user is solving an n-back task, the instructions used to induce different levels of mental workload during the calibration will be based on a n-back task as well, to provide *task similarities* between calibration data and application data. When on the other hand e.g. the task in the calibration paradigm is n-back-based and the classifier trained on this data should then be used to distinguish levels of mental workload in a visual-spatial task, the classifier model may be unable to discriminate workload levels in the visual-spatial task, as the properties of the calibration task and the application task differ too much from another (Arico et al., 2017), even though both tasks induce different levels of the same mental state - which is mental workload.

Hence, it would be desirable to initially design calibration paradigms which induce different stages of the more general neural correlates of the mental state which is to be assessed by the BCI system – not those specific to an application, task or context. That way, the resulting classifier model could be applied to a variety of tasks where the mental state of interest can vary in the user and should be assessed by the passive BCI system. Such a classifier model would be *task-independent*. A mental state can be elicited within different situations and tasks which have a specific context and content, but the evoked neural activity would be the same for the underlying, general mental state of interest. A task-independent classifier trained on data from a calibration paradigm, which produces examples of brain activity highly related to the more general features of the mental state of interest and not to application- or task-dependent features, would enable classifier calibration for a variety of contexts and applications without redesigning a calibration paradigm. This would save time, as the same calibration paradigm could be used to train a classifier model, which can be applied to different tasks and applications. The more general a task-independent classifier model can represent the features of a specific mental state, the more diverse BCI applications and tasks it can be applied to, as a user could complete the calibration paradigm once and use the obtained classifier model for support in different applications and tasks. It should be noted,

that - as with other challenges for passive BCI applications - real-world applications which cannot rely on well controlled environmental properties like laboratory studies do induce additional factors (other mental states, noise etc.) for classifier application across contexts and tasks as situations are generally more complex in real-world scenarios, which may decrease classifier accuracy. A general overview of mental workload related studies investigating cross-task classification to achieve task-independence of the classifier model is given in Gerjets et al. (2014), who conclude, that so far no successful cross-task classification of mental workload could be demonstrated for a classifier.

A well-performing task-independent classifier for mental workload assessment was recently developed and tested by Krol et al. (2016). This classifier model is calibrated using data collected from a subject solving arithmetic exercises that induce brain activity associated with general mental workload. Although the classifier model is taught to distinguish between specific examples of mental workload derived from the arithmetic task, the classifier is meant to assess mental workload in general across multiple tasks. To test this, the classifier has previously been employed for mental workload classification in tasks from other domains, where it was shown that the classifier can correctly distinguish between phases of low and high mental workload, not just in arithmetic tasks but also in anagram solving, backward span tasks, mental rotation, and complex verbal description understanding (Krol et al., 2016; Zhang et al., 2018). The classifier model is therefore assumed to be a task-independent prediction model of mental workload.

In the second study of this work, I investigate whether the prediction model introduced by Krol et al. (2016) may also be successfully applied in a text reading task. To that end, I tested whether the predictive model could be employed for passive BCI-based assessment of mental workload induced by the difficulty and presentation speed of a text during reading. The assessment of mental workload in real time while reading is an essential requirement for the development of neuroadaptive support systems aimed at enhancing the efficiency of reading with digital devices. Further, in case the applied task-independent predictive model can indeed be utilized to identify levels of mental workload imposed by text difficulty and presentation speed during reading, this would extend the scope of tasks the task-independent predictive model developed by Krol et al. (2016) can be applied to and would give further indication of its generalizability to mental workload assessment across tasks.

2.7.7 Novel useful neural correlates for increased versatility in passive BCI applications

There are a number of neural correlates already researched and exploited for passive BCI applications - such as mental workload, error detection, attention, surprise, relaxation assessment, or emotion detection (as detailed in section 2.6.2). However, the scope of potential areas of applications for passive BCI technology could be broadened if even more mental states could be assessed in a user on a single-trial basis through passive BCI technology. There exists a vast corpus of neuroscientific findings from studies which investigated neural correlates of different mental states and processes in human cognition. Since the results reported in most of these studies rely on averaging over trials, it remains unclear if significant effects between classes of the investigated cognitive states imply that the neural correlates of the investigated mental states are detectable on a single-trial basis as well. This potential for detection on *single-trial* basis however is a requirement for a neural correlate of a mental state to be effectively usable in a passive BCI system, where the classifier model bases its decision on the neurophysiological data of only one single trial.

In a study from Appriou et al. (2020), e.g. the single-trial detectability of the mental state of curiosity was investigated. The authors' aim was to learn whether neural correlates of curiosity could be detected and distinguished from other mental states by a passive BCI classifier model on single-trial basis, which would enable identification of the current curiosity level within a person. If states of curiosity were detectable by the passive BCI at a reasonable classification accuracy, the results would be a first step towards adding curiosity to the mental states detectable by a passive BCI and thus enhance the spectrum of applications and contexts in which passive BCIs can be applied. For example learning environments could benefit from curiosity monitoring in students, as educational research has shown that curiosity is a driving factor for the outcome in learning efficiency (Oudeyer et al., 2016).

In their study, Appriou et al. (2020) firstly collect EEG signals during states of curiosity, with the aim to teach a classifier model to discriminate between a user's curious and non-curious states. In a calibration paradigm trivia question and answer chains were used to induce the two relevant states of curiosity and non-curiosity, while EEG data was recorded from 61 electrodes. Best classification accuracies were obtained when a *Filter Bank Tangent Space Classifier* (FBTSC, Appriou et al. (2020)) was employed using a 4 seconds long time-window of data, with an average classification accuracy of 63 percent. Moreover, classifier model analysis indicated that frequency bands from delta to low beta were most discriminative in the classifica-

tion of states of curiosity. The authors conclude that the presented results show that indeed significant classification of different levels of human curiosity are detectable through a passive BCI classifier model. However, it should be further investigated how the underlying neural correlates of curiosity relate to those of mental workload and engagement, if other stimuli can be used in the calibration paradigm to induce levels of curiosity or even if subject-independent classification is possible.

The above described investigation serves as an example of a feasibility study where the single-trial detectability of a neural correlate of a mental state - which to-date is not assessed by passive BCIs - is explored. As detailed above, finding novel neural correlates of mental states that can be detected on a single-trial basis through a BCI, can enhance the versatility and capabilities of NAT, as they open new application contexts and scenarios. In the third study presented in this dissertation (see chapter 5), I investigate such a mental state, whose underlying features of neural correlates have already been researched in the field of neuropsychology but were not assessed on a single-trial basis. The study aims at exploring whether neural correlates of human *moral assessment* - a person's individual judgment of the moral quality of an event or scene that is presented to them - can be measured on a single-trial basis through a passive BCI using EEG data acquisition. If a user's moral assessment could be identified by a passive BCI, this information about a user could be e.g. gathered over time in a user model representing a specific person's moral preferences. Such a user model could provide useful in the interaction between humans and machines in general, or specifically enable a closed-loop adaptation of moral representation in autonomous moral agents toward human moral preferences (see chapter 5). Here the user model containing the moral preferences of a specific user could be adapted and refined over time during human-machine interaction to ideally lead to an alignment in terms of morality between machine and human(s).

Study 1: Subject- and session independent single-trial classification of automation error perception

3.1 Motivation and summary of the study

As one of the grave challenges preventing everyday, real-world applicability of passive BCIs and NAT today is posed by subject- and session-dependence of the trained and applied classifier models (as described in section 2.7.5), this first experiment of my thesis focuses on overcoming this obstacle by investigating the potential for subject- and session-independence of passive BCI classifier models. In particular, this investigation was conducted around an error-detection passive BCI. As described in section 2.6.2 passive BCIs detecting ErrPs as neurophysiological markers from ongoing EEG data measurement, have been the subject of many preceding passive BCI experiments that provided high classification accuracy and even demonstrated the online applicability of the calibrated passive BCI models for error-detection and system adaptation through the classifier's output. Automatic detection and correction of errors through the continuous input generated by a passive BCI model holds a strong potential to improve HMI, as a reduction of errors can reduce a user's workload, improve system performance and increase safety due to a reduction of error-proneness of automated systems. Due to the unobtrusive nature of user-state monitoring through a passive BCI, that does not require additional attention or effort from the user, the additional information channel provided by the error-detection passive BCI does not interfere with the user's primary task or the interaction with the machine.

3.1.1 Experimental paradigm mimicking error-prone HMI and expected ErrP

In the following study, an experimental paradigm was employed, which mimics typical elements of HMI (see section 3.2.1.1). Users are first learning the goal, rules and operations of a game and form individual strategies on how the goal can be reached best and fastest through the interaction with the system. After some time, unexpected errors are induced into the gameplay which prevent the user from executing their formed strategies to reach their goal and require corrective action by the user.

As discussed in section 2.6.2, the elicited ErrP depends on the type of error that is presented and interpreted by the user. In our experimental paradigm the user most likely experiences the induced errors in a way that the machine is not responding to a command in the way intended by the user. This type of error corresponds to interaction or execution errors, hence for the conducted experiment, we anticipate an ErrP that is somewhat similar to those caused by interaction or execution errors.

For the interaction error, which is described as “error[s] of the interface linking human and an external system” (p.1, S. K. Kim and Kirchner (2013)), we would expect an ErrP that shows an ERN around 270 ms after an erroneous event, a subsequent prominent positive peak at around 380 ms, as well as a small second negative peak at around 600 ms as demonstrated by S. K. Kim and Kirchner (2013).

For the execution error, which is observed in others or a system as “inaccurate execution[s] of the selected action” (p.278) which are “mostly surprising events” (p.282) (de Bruijn et al., 2007), we would expect an ErrP that also shows a strong P300 component with a posterior distribution (elicited by the occurrence of an unexpected event), but no ERN (de Bruijn et al., 2007).

Thus, the elicited ErrP that is to be expected should consist of a positivity resembling the P300 component, and could also include an earlier negativity and/or a late negativity.

3.1.2 Investigating subject- and session independence in an error-detection passive BCI

The aim of this study was to investigate the potential and applicability of an error-detection passive BCI whose classifier model is session- and/or subject independent. The benefits for real-world applicability of passive BCIs that would be provided by the development of session- and/or subject independent classifier models - which do require minimal or no calibration time and data - would be immense and were already detailed in section 2.7.5.

The study was structured and conducted in seven parts, each of which aims at investigating distinct research questions. Here I will firstly present an overview of

the different parts of the study which are later detailed in sections 3.2.1 to 3.2.7:

- **Part 1: Neurophysiology of a subject-specific error-detection classifier (see section 3.2.1)**

We first calibrated a classifier model for single-trial error-detection on the calibration data gathered from the experimental design that included correct and erroneous trials. These classifier models were *subject-specific* and we investigated each subject-specific classifier's predictive performance on new data (offline) from the same subject. Further, we aimed to study the neurophysiology of correct and erroneous trials of the classifier calibration data in an ERP analysis with respect to the literature on error perception. Finally, we visualized the underlying neurophysiology that the subject-specific classifier models were basing their decisions on.

Finding that these subject dependent classifier models were mainly operating on data stemming from eye activity, as opposed to cortical sources that a BCI should be based on, we cleaned the EEG data from activity that was generated by eye components, defining new calibration data sets which excluded eye activity.

We also investigated whether a subject-specific classifier model with sufficient accuracy could be calibrated on data from EOG channels only (which would require less preparation time than EEG headset application). We found that these classifier accuracies were low, though not random.

- **Part 2: Calibrating subject-specific error-classifiers on data cleaned from eye activity (see section 3.2.2)**

We again calibrated subject-specific classifier models, but now used the subject-specific data sets that were cleaned from eye activity. We again performed ERP analysis, classifier calibration and classifier model visualization to find out whether the newly generated classifier models now discriminated between erroneous and correct trials on the basis of cortical activity, instead of eye activity. We also investigated whether omitting eye components from the data had an effect on classification accuracies.

We found that the classifier models were indeed basing their decisions on activity from cortical sources and these classifier models still showed high classification accuracy.

- **Part 3: Session-independence - Applying subject-specific classifier models from a first recording session to new subject-specific testdata from a second recording session (see section 3.2.3)**

To investigate session-independence, we tested whether the subject-specific classifier models that were generated in part 2 could be applied to new data of the same person from a new recording session from a different day. We compared the performance of the classifier model on the new data to its performance on testdata from the first recording session. The results showed no significant difference between the application of subject-specific classifiers to testdata from the same or a different recording session – hence the subject-specific error-detection classifiers are indeed session-independent, as they can be transferred from one recording session to another with similarly high classification accuracies.

- **Part 4: Session independence - Training a group classifier on calibration data from the first recording session and applying it to new subject-specific testdata from a second recording session (see section 3.2.4)**

For further investigation regarding session-independence, we examined whether a session independent classifier can be trained on data from multiple subjects as well. Hence, we trained the error-detection classifier not only on data of one specific subject, but trained a group classifier model on all calibration data of the first recording session and then applied these group classifier to the subject-specific testdata of the second recording session. We compared the group classifier's performance to that of the subject-specific but session-independent classifier models and further investigated whether the group classifier is working on the same cortical sources for classification as the subject-specific classifier models from part 2.

We found no difference between classifier performance of the subject-specific and the group classifier, and the classifier's visualization also showed strong similarities. The results suggested, that we can train one single classifier model on calibration data from a group of subjects and yield similar prediction error rates as with the specifically trained models. Across subjects, this session-independent group classifier seems to work on the same discriminative neurophysiological features as the specifically trained classifier models – indicating that neural responses to error perception are highly similar across subjects.

- **Part 5: Subject-independence - Applying subject-specific and group classifier models to subject-specific testdata from a different subject (see section 3.2.5)**

The group classifier calibrated in part 4 was partially informed about data from the specific subject whose new testdata we applied it to (as the group classifier is trained on calibration data from all subjects - including the test

subject) and therefore still subject-specific.

Thus, we now aimed to investigate subject-independence of the error-detection classifier. To that end we applied subject-specific classifier models to test-data from other individual subjects. Additionally, we trained group classifiers, excluding a specific subject's calibration data from the group classifier calibration, and then applied this leave-one-subject-out group classifier to the subject-specific testdata of the subject, who was left out of the calibration dataset.

Results showed that both, the subject-specific and group classifier models perform worse on new data from a different subject, than the subject-specific classifier model, calibrated on data from the same subject. Classification and application accuracies were significantly better in the group classifier than in the subject-specific classifier trained on data from only one different subject. Even though the application accuracy of the group classifier was on average around 6 percent worse than accuracies of the subject-specific classifiers, this loss in classification performance should be deemed acceptable, considering that the application of a group classifier can be employed without collecting any calibration data from the specific user, enabling a zero-calibration plug-and-play BCI classifier, which would be necessary for feasible real-world applications of (error-detection) passive BCIs (as discussed in 2.7.5).

- **Part 6: Reducing the number of EEG channels used and omitting channels recording eye activity (see section 3.2.6)**

Reducing the number of needed electrodes is crucial for real-world application of passive BCIs, as users want to spend minimal time setting up and preparing hardware before they can start using the BCI application for support and adaptation in HMI. Thus, we tested whether the number of electrodes of the applied setup can be reduced, without decreasing classifier performance of the error-detection group classifier.

Using the calibration data from the first two recording sessions for classifier training, we first reduced the initial 50 electrode setup down to 32 and then to 24 electrodes. Classifier calibration performance of subject-specific classifiers showed that there was a statistically significant but practically dismissible (on average 1 percent) decrease in calibration accuracy from the 50 electrode setup to both, the 32 and the 24 electrode setup. There was no difference found between the 32 and the 24 channel setup. Calibration accuracy of the group classifier model increased slightly from 50 electrodes to both, the 32 and 24 electrode setups. These results indicated that we could reduce the number of electrodes for the group classifier calibration down to 24 channels.

- **Part 7: Session- and Subject independence - Application of the group classifier to data from new subjects from recording session 3 (see section 3.2.7)**

In the last part of this study, we worked towards combining session- and subject-independence of the error-detection group classifier by transferring it to new data from a third recording session with new subjects. We calibrated the group classifier on data from 24 EEG channels from the first two recording sessions and explored its application performance in comparison to subject-specific classifiers trained on subjects' data from a new third recording session. To examine the group classifier's performance in a real-world scenario, subjects in this new recording session experienced online adaptation and correction through the error-detection BCI.

Results showed, that there was no difference in the estimated classification accuracy of the subject-specific classifier models and the newly defined group classifier. In the online error-correction scenario, application accuracy of the group classifier was on average at 77 percent. These results showed, that for automatic error-detection we can train a reliable session- and subject-independent group classifier, on calibration data of multiple subjects, with a reduced electrode number of 24 channels. Application accuracies of this group classifier remain stable during online application, where the classifier is used for real-time detection of error potentials and automatic error correction while the user performs their task.

After this overview of the structure of this study, in the following section (3.2) the methods, results and discussion of the seven distinct parts of this study are presented in detail.

3.2 Study parts 1 to 7: Methods, results and discussion

3.2.1 Part 1: Neurophysiology of a subject-specific error-detection classifier

In this first part of the study we calibrated a standard subject-specific classifier model for single-trial error detection through feature extraction with a windowed-means approach (Blankertz et al., 2011). We investigated the classifier's predictive performance on new data and visualised the underlying neurophysiology this classifier bases its decisions on.

3.2.1.1 Methods

Participants In this first experimental data recording session, data from 14 subjects was recorded (8 male, 6 female, age: 21 to 30 years). For all recorded data in this study, data recording has been approved by the ethical committee of the Technical University Berlin and informed consent was given by all subjects.

Apparatus EEG data was recorded from 50 electrode sites with a sampling rate of 1000 Hz. Vertical and horizontal electrooculogram (EOG) was additionally recorded from 4 electrodes. We used Brain Products EasyCap electrodes, with electrode placements according to the 10/20 system and the signals were amplified with a BrainAmp DC system (Brain Products GmbH, Gilching, Germany). The reference was the nasion and the ground electrode at Fpz. All electrode impedances were kept below 20k Ω .

Experimental paradigm As an experimental task, a game was designed which simulated a realistic HMI scenario, as there was a clear task, rules and a defined set of actions for the subject to perform in interaction with a machine.

In this experimental paradigm, the *Rotation-Left-Right game* (RLR-game, Zander et al. (2009), see figure 3.1), it is the objective of the user to rotate one of two letters (either "R" or "L"), that is shown on top of a circle, until the letter has reached an indicated target position. As controls for the rotation of the letter, a left key press is used for rotating the "L" counterclockwise, while the "R" is to be rotated by a right key press in clockwise direction. Every second, the color of the letter of the current round is changing. The color indicates the degree of the rotation that is elicited if the key gets pressed – if the color of the letter is red, it will be rotated by 90 degrees, and if it is green the rotation will be by 30 degrees. In every one second long phase only one key press is possible.

The goal of each round of the game is to position the letter at the indicated goal position as fast as possible without rotating it further than the exact target position (see figures 3.1 and 3.2). As the user is free to press the key in every one second long phase or not, they can choose and develop different strategies to reach the target position (e.g. waiting for red color of the letter to only make big rotations of 90 degrees and omitting the green color phases, or choosing to press the key every second in the beginning of a round and in the end wait for the right color to rotate the letter into its target position).

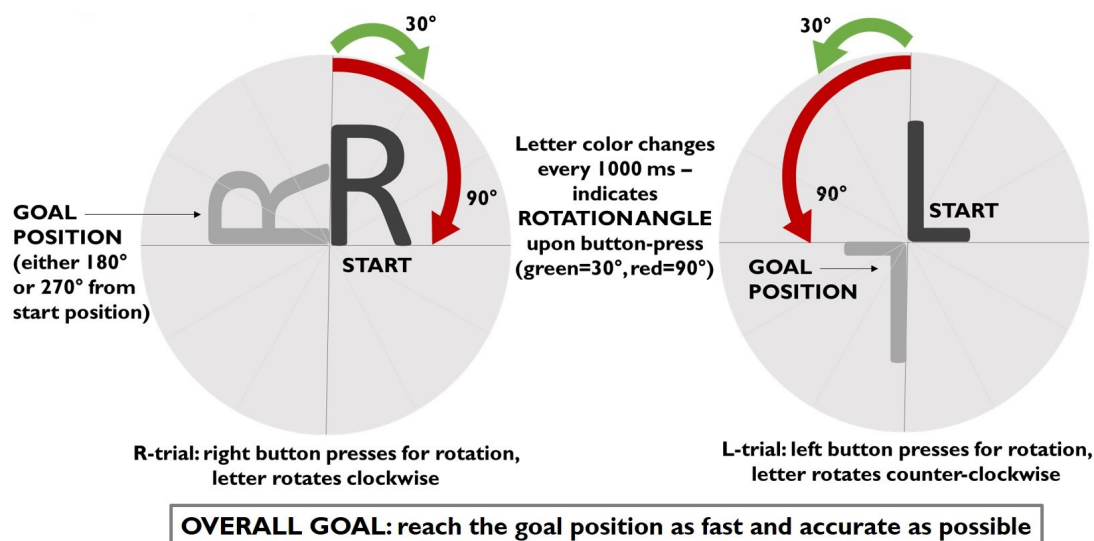


Fig. 3.1 Experimental paradigm - the RLR-game used in the three different data recording sessions

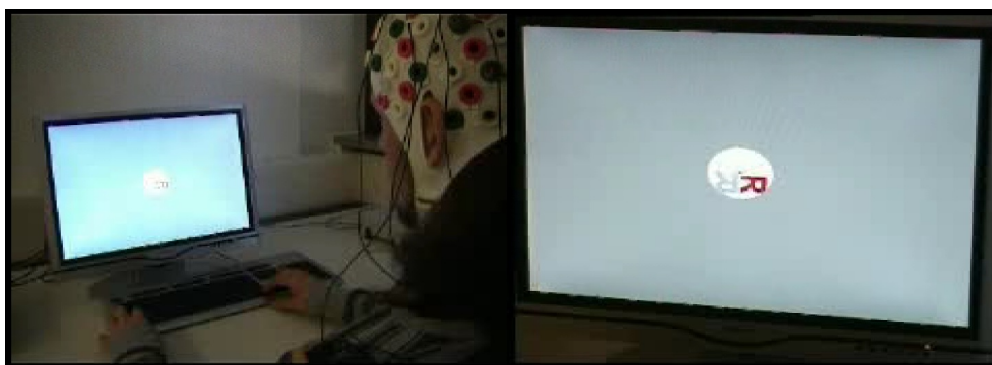


Fig. 3.2 Experimental setup for playing the RLR-game

In our experimental setting there are different phases of the game: in the first phase the user simply plays the game as detailed above. In a second phase, errors are induced randomly at a total of 30 percent of key presses into the game: the rotation in these trials is smaller than anticipated by the user, as red letters will only

turn by 30 degrees (instead of by 90 degrees) and green letters will not turn at all (instead of by 30 degrees) – see figures 3.3 and 3.4. In some parts of the three recording sessions of this experiment in which the user plays the RLR-game, they play against an opponent to increase motivation to perform well in the game. When competing against an opponent, the current score of the game is displayed to the user after each round where the point goes to the player who reached the target first.

The errors induced into the RLR-game correspond to *execution or interaction errors* in the automation in an HMI context, as there is a mismatch between the game's response to the action of the button press and the anticipated result of the button press formed in the user. The induced errors also elicit frustration in the user as they prevent them from following a built up strategy for playing the RLR-game.

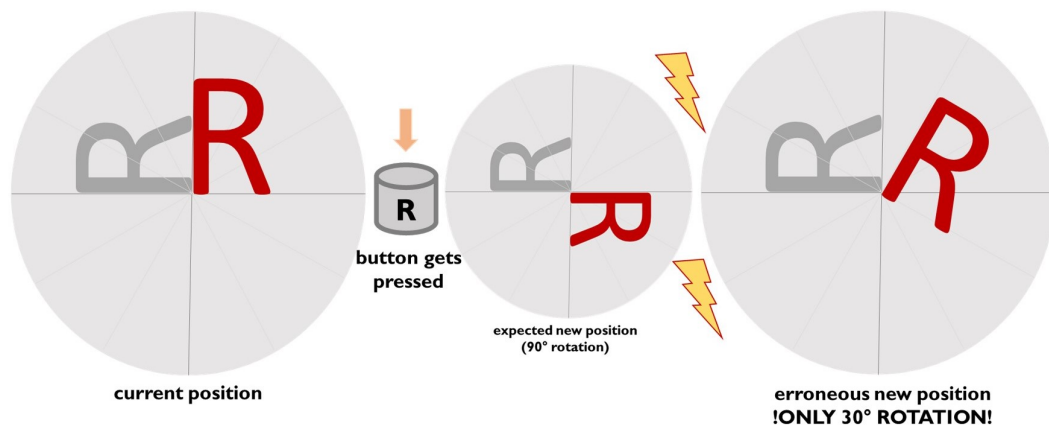


Fig. 3.3 Errors in the RLR-game: errors occurring at red letter, when 90 degree rotation is expected. Errors occurred at 30 percent of button presses.

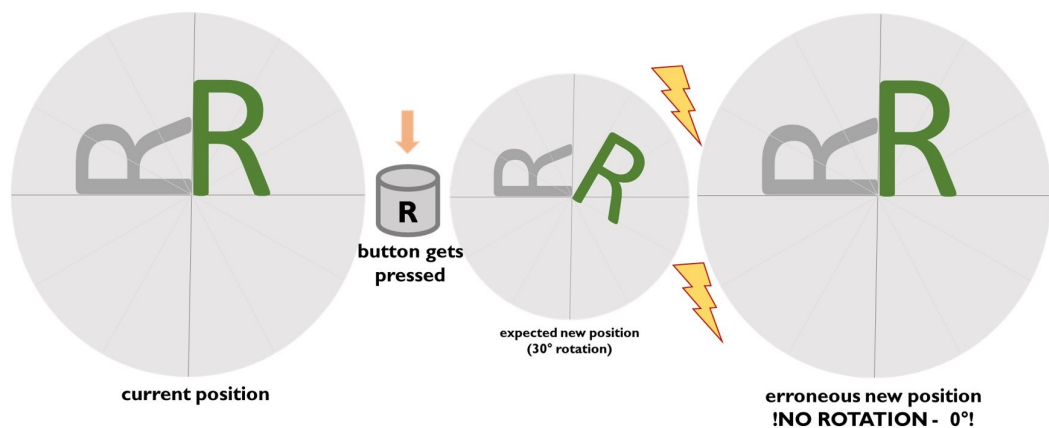


Fig. 3.4 Errors in the RLR-game: errors occurring at green letter, when 30 degree rotation is expected. Errors occurred at 30 percent of button presses.

Procedure Subjects in the user training got a chance to play the game without an opponent while learning the rules, sitting alone in a closed recording booth. In a first recording block they played the RLR-game without any errors induced into game-play. In a second recording block, subjects played the RLR-game again, this time machine errors were introduced at a rate of 30 percent of trials.

In a third recording block subjects played the RLR-game against a trained opponent, again without any machine errors induced. During this block, the subject's task remained unchanged to blocks one and two, but the introduced opponent was sitting nearby, visible to the subject. In a following fourth recording block, the subject played against the opponent while machine errors were induced into the RLR-game, again at a rate of 30 percent. In each of the four recording blocks, approximately 40 rounds of the RLR-game were played.

3.2.1.2 Data Analysis

We analysed the data from this recording session by firstly plotting and examining event-related potentials (ERP) from trials with and without machine errors. We also trained subject-specific classifier models on these two trial classes and analysed which cortical areas contributed to these classifiers' decisions.

All data recorded in this study included a known lag in timing of 100 ms due to the recording setup. This lag was corrected for in all recorded datasets.

Event-related potentials We used the first two datasets for plotting grand average ERPs. We filtered the data, applying a lowpass filter of 40 Hz and a highpass filter of 0.5 Hz. Data was rereferenced to the common average and epochs of 1.2 seconds length were extracted starting 200 ms before stimulus onset of erroneous and correct rotations. A baseline correction was performed using the 200 ms before stimulus onset. Grand average ERPs were generated from all trials of all subjects for erroneous and correct rotations. These grand averages were plotted for electrode sites Fz, Cz, and Pz. The respective ERP difference curves (error minus no error) between the two trial classes were also plotted, including p-values resulting from per sample permutation tests. A permutation test is a non-parametric test and therefore does not make assumptions about distributions, providing a solution to the multiple comparisons problem (Maris and Oostenveld, 2007). To compare two conditions, condition labels are shuffled n-times and the means of obtained distributions are compared to the mean of the distribution with original labels. If data of both conditions came from the same distribution, which is the null hypothesis, means of distributions with shuffled labels should not differ significantly from the mean of the original distribution.

Classifier calibration For each of the 14 subjects, we appended the first two recording blocks of the respective subject into one new dataset for calibrating the classifier on these data. Before appending the two datasets for each participant, we removed EOG channels, reducing the number of channels from 54 to 50. We then set a lowpass filter of 200 Hz, down-sampled the data to 100 Hz and rereferenced to the common average. Offline classification was performed using the open toolbox BCILAB (Kothe and Makeig, 2013) where features were extracted through the built-in windowed-means approach (Blankertz et al., 2011). In this approach, first a bandpass filter from 0.1 to 15 Hz was applied to the data. Epochs of 700 ms length were extracted starting at stimulus onset and ending at 700 ms. For 7 consecutive time windows of 50 ms length between 150 and 500 ms after stimulus onset, the average amplitude was calculated. Classification was performed through LDA. A (5x5)-fold cross-validation was performed to generate an estimate of the model's online reliability.

Classifier visualisation From every subject's individual classifier model, a neuroscientifically interpretable pattern was generated for every time-window. These patterns indicate which channel activity was relevant for the classifier's decision to separate the classes in the specific time-window. That way, information used by the classifier to differentiate between classes was revealed and applied for validation of the classification approach. To that end, a normalization of the obtained subject-specific patterns over all time-windows was performed. Then, these patterns were averaged across all subjects in every time-window, yielding seven grand average patterns.

For identifying and visualising sources relevant for classification, the LDA-based classifiers were analyzed, by transferring the LDA filter weights from the sensor level to ICs in the source space by using the source localisation tool (Krol, Mousavi, et al., 2018) as described in 2.5.4. We generated weighted dipole density plots showing the relevance of cortical areas to the windowed-means classifier in a 3-dimensional visualisation. The resulting relevant ICs serve as an indication to identify cortical areas that are sensitive to error-perception.

3.2.1.3 Results

In this section, results of the ERP analysis, classifier calibration and application to testdata are presented, as well as classifier visualisation from this first data recording session.

Event-related potentials The grand average ERPs of correct and erroneous trials on electrode sites Fz, Cz, and Pz are depicted in figure 3.6. The respective ERP dif-

ference curves (error minus no error) between the two trial classes are shown in figure 3.5. To further analyse the differences between the two trial classes, for every sample in the ERP a permutation test with 5000 permutations was performed (we chose this amount of permutations to obtain an accuracy better than 0.001 of the p value, with a confidence interval of 99 percent (Ojala and Garriga, 2010)). P-values of samples with significant permutation test outcomes are indicated as gray bars in the background.

The results indicate that there are no significant differences between correct and erroneous ERP curves at electrode Fz, though a small negativity around 230 ms can be observed followed by a positivity that has its strongest peak at around 340 ms.

At electrode Cz no early negativity can be observed in the ERP difference plot, but a strongly pronounced positivity for erroneous trials with a peak at around 310 ms is present. According to the p-values resulting from per sample permutation tests this positivity is also highly significant between 250-420 ms.

At electrode site Pz again no early negativity can be observed in the difference plot, but a positivity for erroneous trials with a peak at around 310 ms is present. The p-values resulting from per sample permutation tests are significant between 300-370 ms for this positivity.

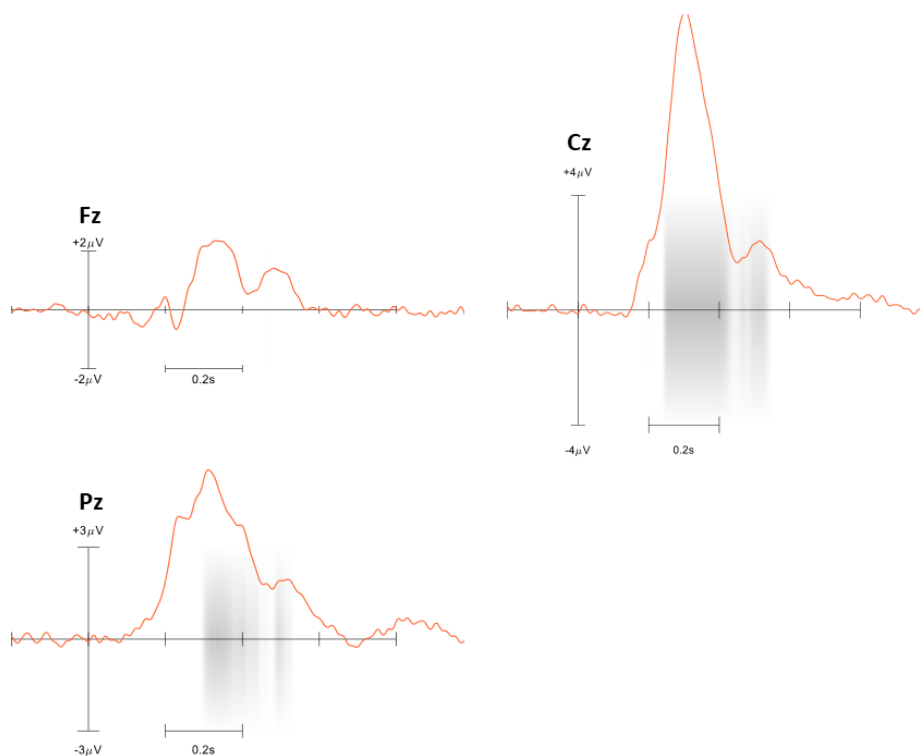


Fig. 3.5 Grand average difference curve of the ERP (error minus no error) of the first recording session at electrode sites Fz, Cz and Pz. Gray bars behind the plot indicate significant p-values from the permutation tests that were performed for each sample.

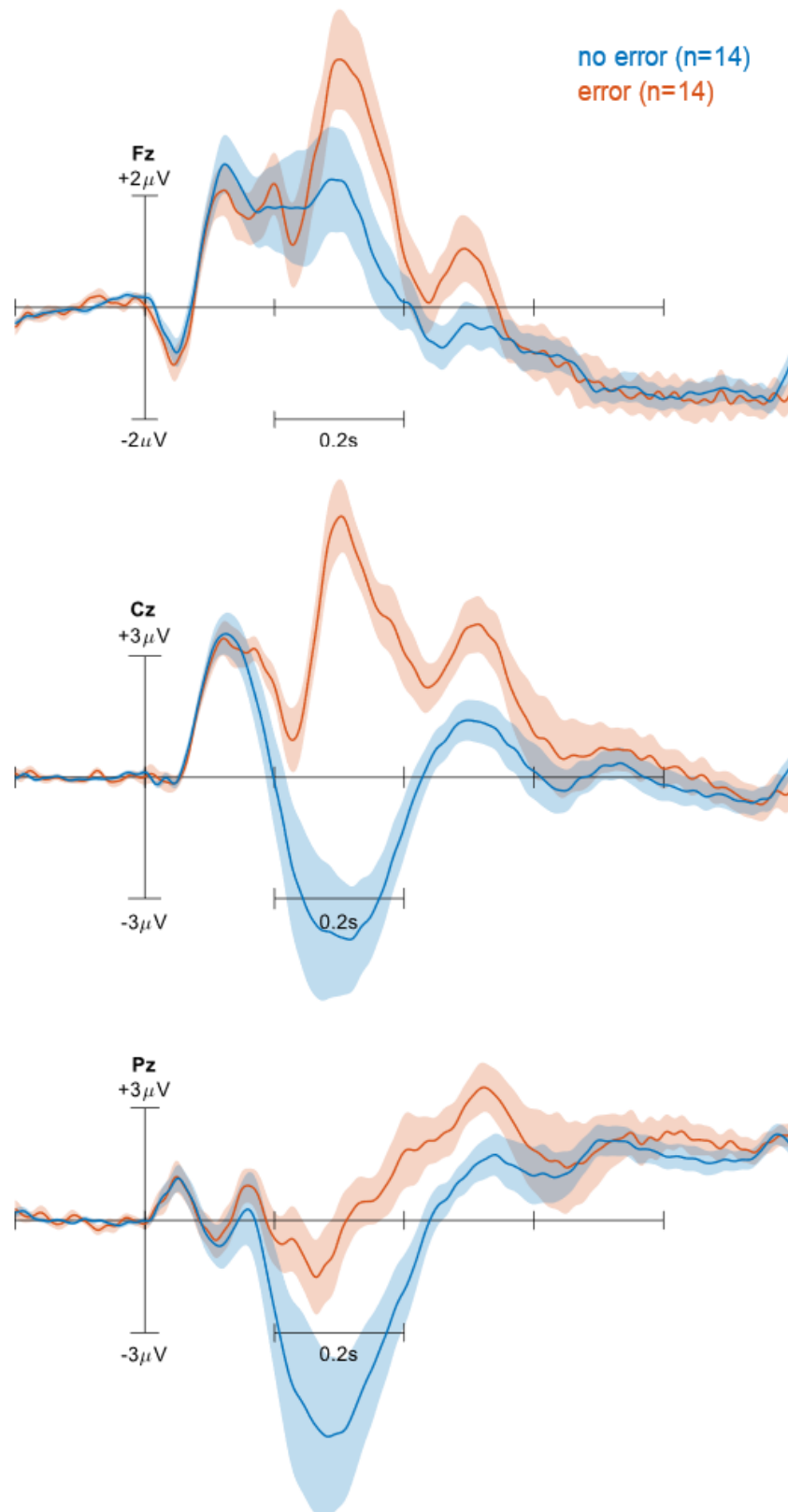


Fig. 3.6 Grand average ERPs ($n=14$) of the first recording session. Curves show error and no error trials from electrode sites Fz, Cz and Pz with standard deviations.

Classifier calibration For all 14 participants the obtained classification error rates are shown in table 3.1. The average classifier calibration error rate was 8.73 percent (SD=2.61 percent).

Table 3.1 Classifier calibration error rates in percent with standard deviations of subject-specific classifiers from recording session 1 (data includes eye activity)

subject	calibration error	SD
1	11.61	0.83
2	12.77	2.42
3	4.87	2.47
4	5.06	1.61
5	11.06	3.59
6	15.91	3.77
7	7.94	2.65
8	6.65	1.74
9	5.79	2.79
10	9.01	2.44
11	17.88	5.60
12	5.32	3.23
13	6.50	1.45
14	1.86	1.98
average	8.73	2.61

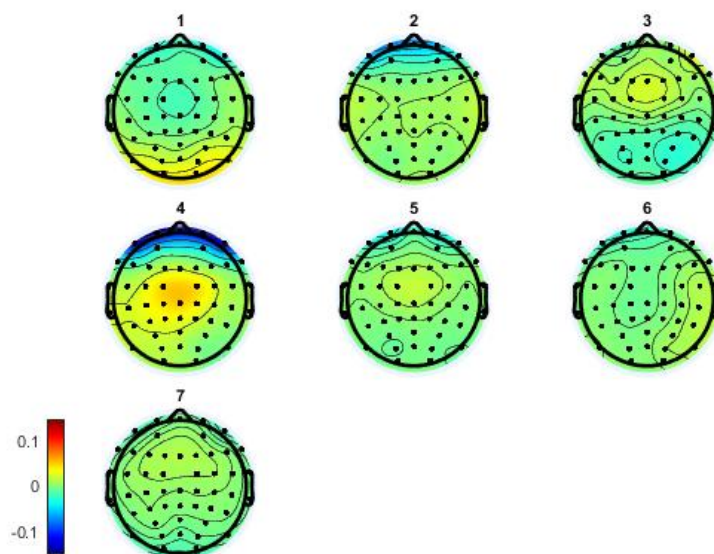


Fig. 3.7 Grand average patterns from each 50 ms long time window from 150 to 500 ms of the classifier models of all 14 subjects

Classifier visualisation The normalized grand average patterns from each time window from 150 to 500 ms of the classifier models of all 14 subjects are shown in figure 3.7. The patterns show that the classifiers focus most weights on frontal areas, with most weights on the time window 300-350 ms. This weight distribution indicates, that the classifier models base their decisions mostly on eye activity instead of activity coming from cortical sources in the brain.

Further analysis with the source localisation tool showed that indeed sources used for the classifiers' decisions are mainly focused around the eye area. Source localisation for all seven 50 ms time windows from 150 to 500 ms are depicted in figures 3.8 to 3.14.

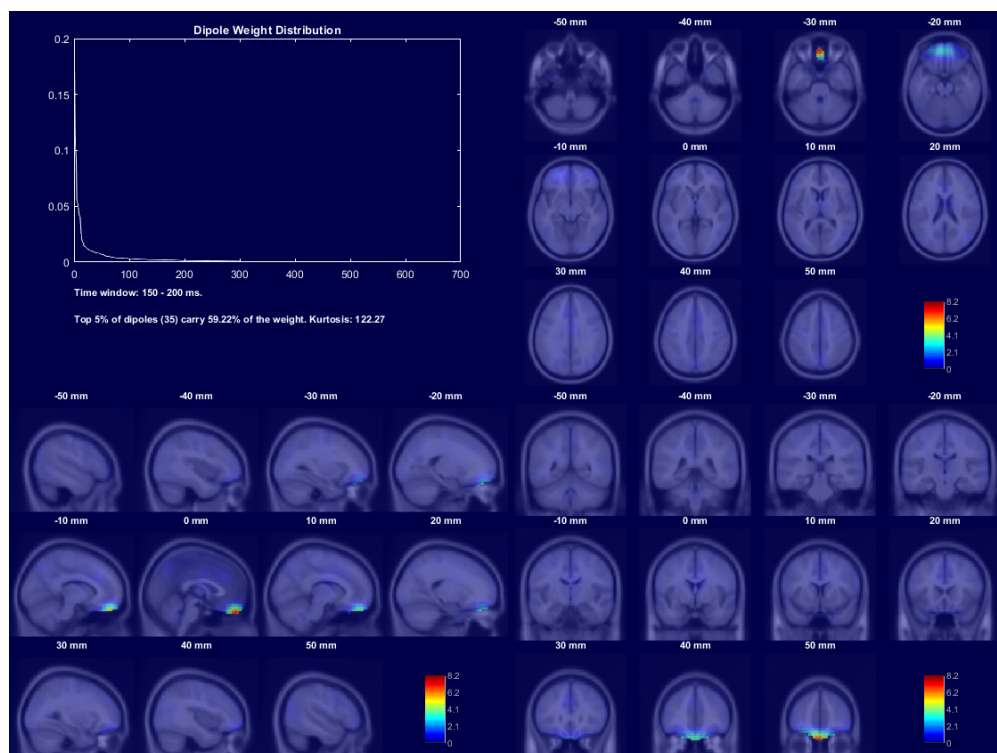


Fig. 3.8 Visualisation of the subject-specific classifier including data from eye components in time window 150-200 ms: Weighted dipole density plots showing the relevance of cortical areas to the windowed means classifier. Slices are labelled with their corresponding MNI coordinates. Top left: sorted dipole weight distribution.

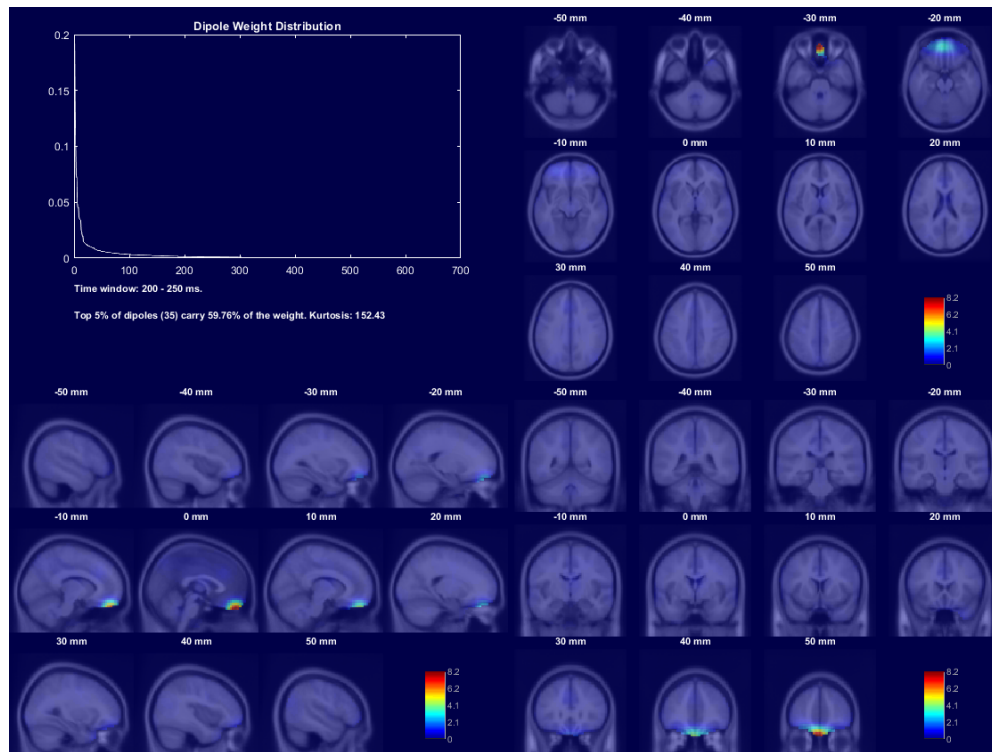


Fig. 3.9 Visualisation of the subject-specific classifier including data from eye components in time window 200-250 ms.

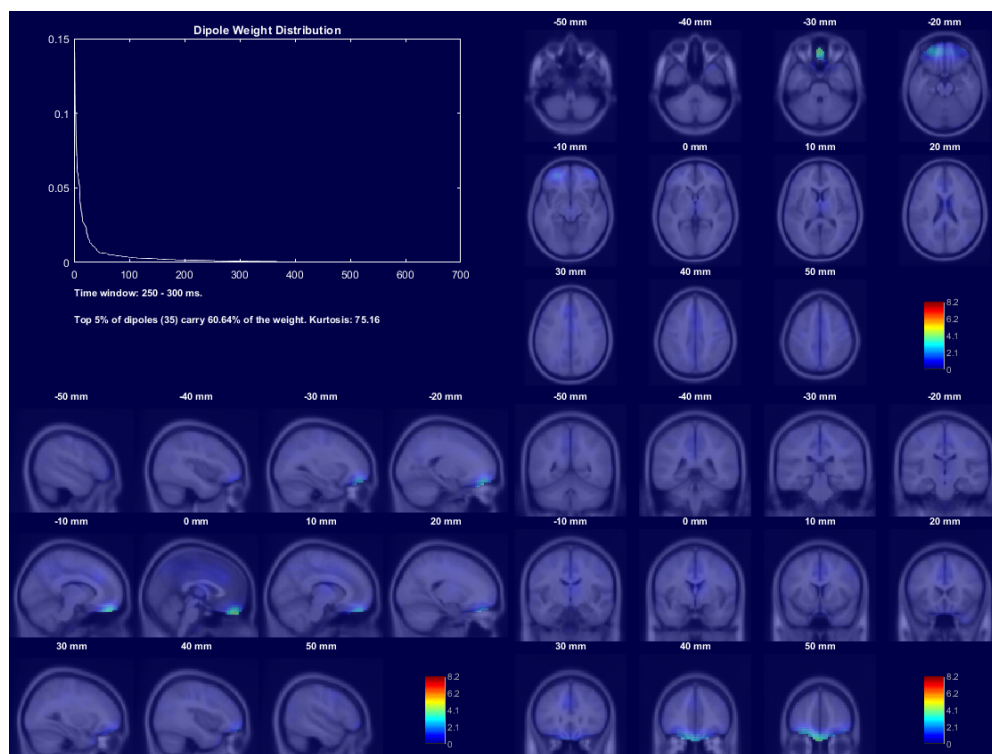


Fig. 3.10 Visualisation of the subject-specific classifier including data from eye components in time window 250-300 ms.

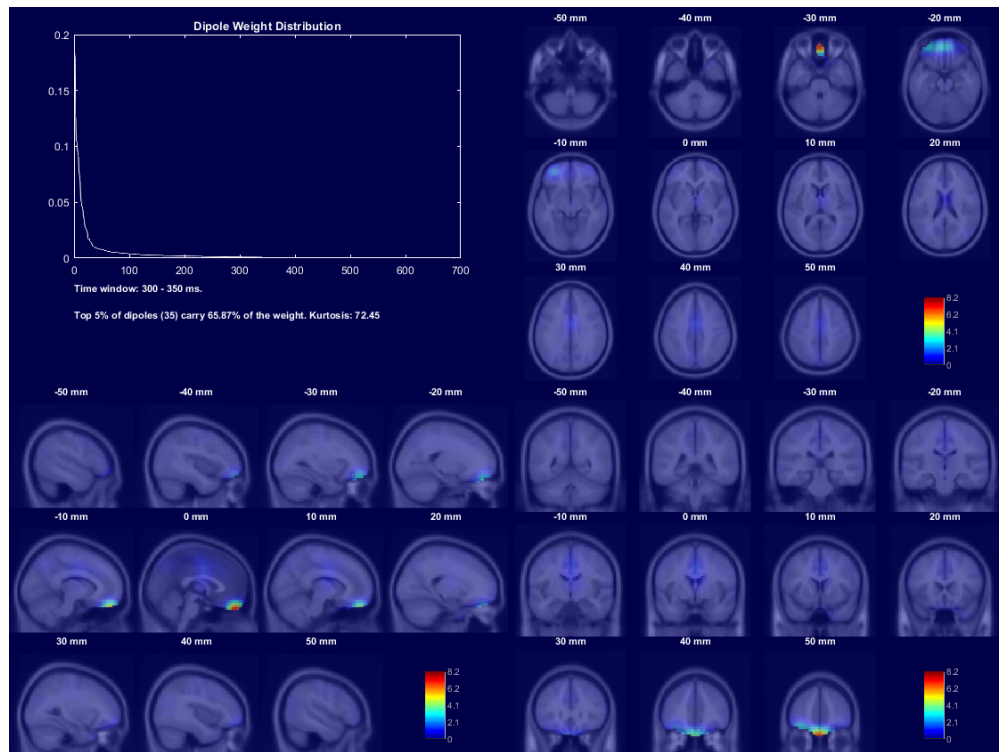


Fig. 3.11 Visualisation of the subject-specific classifier including data from eye components in time window 300-350 ms.

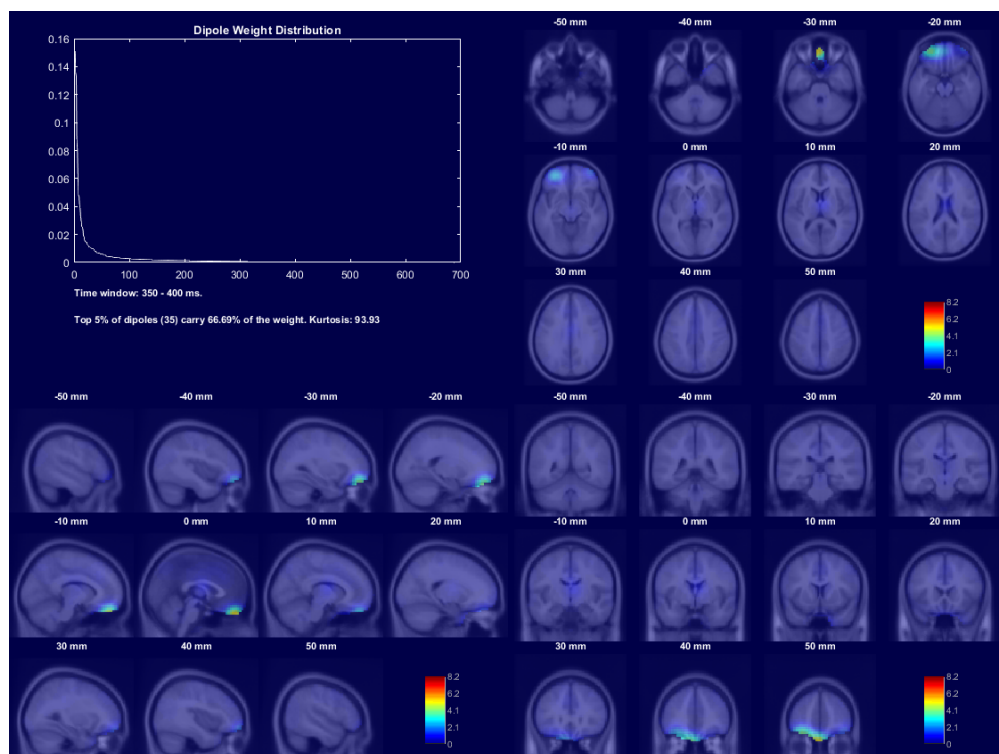


Fig. 3.12 Visualisation of the subject-specific classifier including data from eye components in time window 350-400 ms.

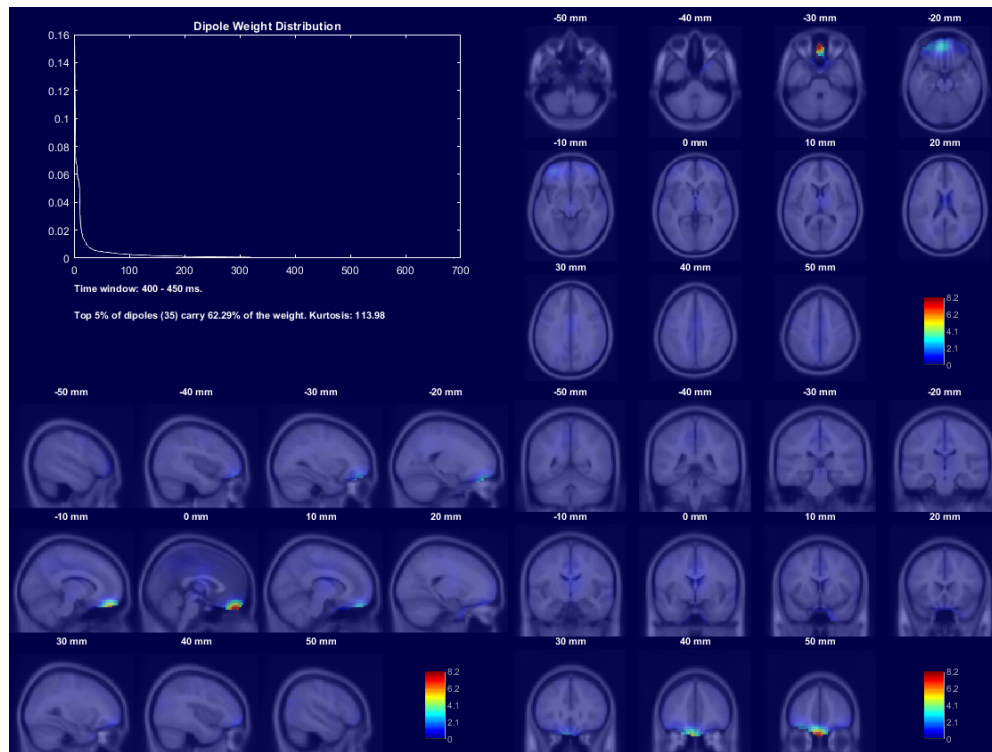


Fig. 3.13 Visualisation of the subject-specific classifier including data from eye components in time window 400-450 ms.

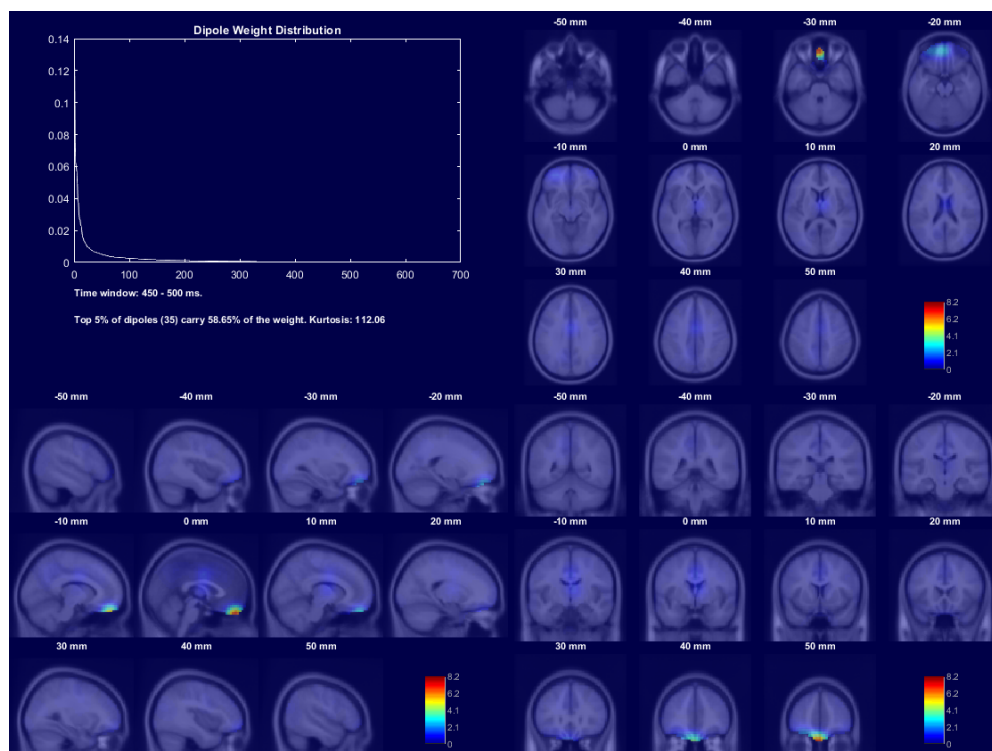


Fig. 3.14 Visualisation of the subject-specific classifier including data from eye components in time window 450-500 ms.

Classifier application to testdata We applied the calibrated subject-specific classifier models to subject-specific testdata from the fourth recording block from this first recording session. The resulting average prediction error was at 12.18 percent. The obtained prediction errors for each subject are shown in table 3.2.

Table 3.2 Classifier prediction errors rates in percent from application of the subject-specific classifiers to testdata from recording session 1 (data includes eye activity)

subject	prediction error
1	15.00
2	11.67
3	8.66
4	10.90
5	8.89
6	17.33
7	17.24
8	11.08
9	8.81
10	8.68
11	19.48
12	14.04
13	12.60
14	6.16
average	12.18

Further analysis of subject-specific classification on EOG channels As we saw in the above presented results that subject-specific classifier models are predominantly putting relevance weights for classification on frontal electrodes at areas around the eyes, we also calibrated subject-specific classifier models for each subject, taking only data from EOG channels into account. Here, we wanted to clarify if recording on only the four EOG channels would result in similar classifier calibration and application results, as the subject-specific classifier models that were trained on EEG data from all 50 channels. If similar results could be achieved, recording from 50 plus channels would become unnecessarily cumbersome for classifier application - and instead of working with a BCI one could work with an easier and faster to handle *eye-computer interface* (ECI) with just the four EOG channels for data acquisition.

For this analysis, we used the same classifier calibration- and testdata as before, keeping only the four EOG channels in the data and omitting the other 50 EEG channels. Data was then pre-processed in the same way as described above, only it was not rereferenced. We then calibrated subject-specific classifier models on the EOG data (with the same approach as described above) and applied the obtained models to the respective subject's testdata. Results showed, that the average calibration error when training the classifier models on EOG channels was at 36.07 percent (SD = 5.23 percent), see table 3.3. The classifier application error rates of these subject-specific classifier models to EOG data was on average at 39.87 percent, see table 3.4.

Table 3.3 Classifier calibration error rates in percent with standard deviations of the subject-specific classifiers trained on EOG channels

subject	calibration error	SD
1	52.46	11.43
2	31.62	5.84
3	30.18	4.58
4	25.93	5.67
5	46.35	6.74
6	44.38	3.91
7	31.03	6.81
8	29.93	1.99
9	41.94	3.57
10	33.45	6.35
11	29.64	1.84
12	29.58	4.50
13	42.81	4.14
14	35.67	5.86
average	36.07	5.23

Table 3.4 Classifier prediction errors rates in percent from application of the subject-specific classifiers trained on EOG channels to testdata

subject	prediction error
1	50.33
2	29.65
3	48.74
4	34.62
5	38.73
6	42.39
7	51.41
8	27.85
9	29.83
10	42.51
11	30.34
12	42.81
13	55.34
14	33.65
average	39.87

3.2.1.4 Discussion

ERP analysis of the EEG data recorded from subjects completing the RLR-Game paradigm showed that classic components of ErrPs are partially present in our data. On central and parietal sites, we find no ERN, but a positivity peaking around 310 ms is present for erroneous trials. This component is not present in trials without induced errors. Hence, the setup of trials in the RLR-Game paradigm indeed induces the intended perception of execution error, which is reflected in the ErrP components of the grand average ERP curves.

Calibration results for subject-specific error-classifier models from EEG data show a low error rate, which is consistent over subjects. However, further neurophysiological visualisation of information used by the classifier to discriminate between classes, revealed that the generated classifier models mainly base their decisions on frontal electrodes and sources, that are mainly reflective of eye activity. As recording from only four EOG channels would be a more convenient in real-world applications, we tried to calibrate and test classifier models following the same approach but by only taking data from EOG channels into account. The results showed, that eye activity holds some useful information for the error-detection classifier and classification is not fully random, but the calibration and classifier application accuracies, which on average are only at 60 percent, are still too low for meaningful and applicable usage for real-world classification. Aside from that, we aim to train a *brain*-computer interface, which bases its classifier's decisions not on eye activity, but on brain activity with cortical origin. For the next step in this study we therefore decided to compute an ICA solution of the data from this recording session to remove activity in the EEG recordings which originates from eye components. On these data, cleaned from eye activity, we then calibrated and tested new subject-specific classifier models (again following the already described classification approach).

3.2.2 Part 2: Calibrating subject-specific error-classifiers on data cleaned from eye activity

In this second part of the study we calibrated subject-specific error-classifier models again, but this time calibration data was cleaned from eye activity prior to classification by computing an ICA solution for each subject and removing activity of ICs reflecting eye activity from the respective subject's EEG data. We performed ERP analysis, classifier calibration and classifier model visualisation again, as in study part 1 (3.2.1.2), to find out whether the newly generated classifier models now discriminated between erroneous and correct trials on the basis of cortical activity, instead of eye activity.

3.2.2.1 Methods

Participants In this part of the study, the same data from the same subjects (see section 3.2.1.1) as in study part 1 was used.

Data Analysis For each subject, we cleaned their data from eye activity by subtracting those parts of their EEG data which were modulated by eye activity from the rest of the EEG dataset. To identify which independent components of each subject's ICA solution of the data were reflective of eye activity, we used EEGLAB's *ICLabel* (Pion-Tonachini et al., 2019) plugin which is an automatic EEG independent component classifier. We regarded ICs that were classified as eyes with more than 50 percent probability as eye components. These eye components were subtracted from the specific subject's EEG data - resulting in a new dataset cleaned from eye activity. The datasets were then subjected to ERP analysis, classifier calibration and classifier visualisation again in the exact same way as in study part 1 (see section 3.2.1.2).

3.2.2.2 Results

In this section, we present the results of ERP analysis, classifier calibration and application, as well as classifier visualisation from EEG data of the first recording session which was cleaned from eye activity using ICA.

Event-related potentials The grand average ERPs on electrode sites Fz, Cz, and Pz from the new datasets cleaned from eye activity are depicted in figure 3.16. The respective ERP difference curves (error minus no error) between the two trial classes are shown in figure 3.15. We again analysed the differences between the two trial classes by performing a permutation test for every sample in the ERP with 5000

permutations. P-values of samples with significant permutation test outcomes are indicated as gray bars in the background.

Morphology and time course of the ERP at Fz are similar to the ERP from study part 1, but the difference curve shows that the positivity, now peaking at 300 ms, is more pronounced after activity of eye components was removed from the EEG data. Permutation tests also yielded significance for this positivity between 250-400 ms.

At electrode Cz a very small, not significant early negativity with a peak at around 150 ms is present. As in study part 1, again a strongly pronounced positivity for erroneous trials with a peak at around 310 ms can be observed. Similar to part 1 of this study, this positivity is again highly significant between 260-420 ms. In the difference curve also another broad significant positivity can be observed from 500 to 680 ms.

At electrode site Pz the difference curve strongly differs from the first part of the study after activity of eye components was removed from EEG data: now a delayed early negativity is present with a peak at around 270 ms. This negativity is significant between 240 and 300 ms. This negativity is followed by a broad positivity that is significant from 490 to 650 ms.

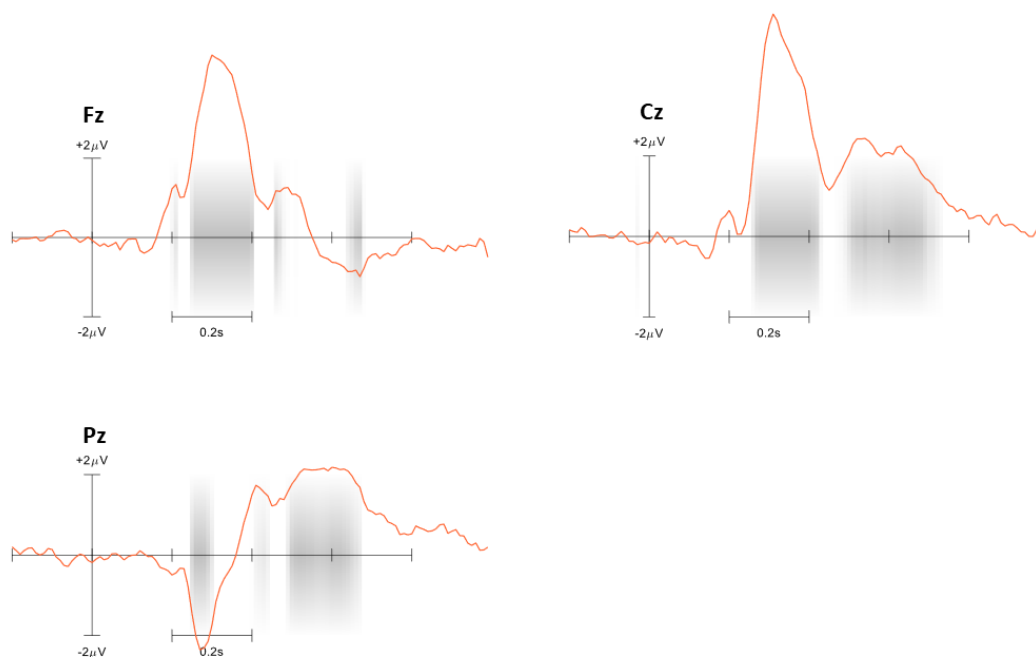


Fig. 3.15 Grand average difference curves of the ERP (error minus no error) of data cleaned from eye activity through ICA of the first recording session at electrode sites Fz, Cz and Pz. Gray bars behind the plot indicate significant p-values from the permutation tests that were performed for each sample.

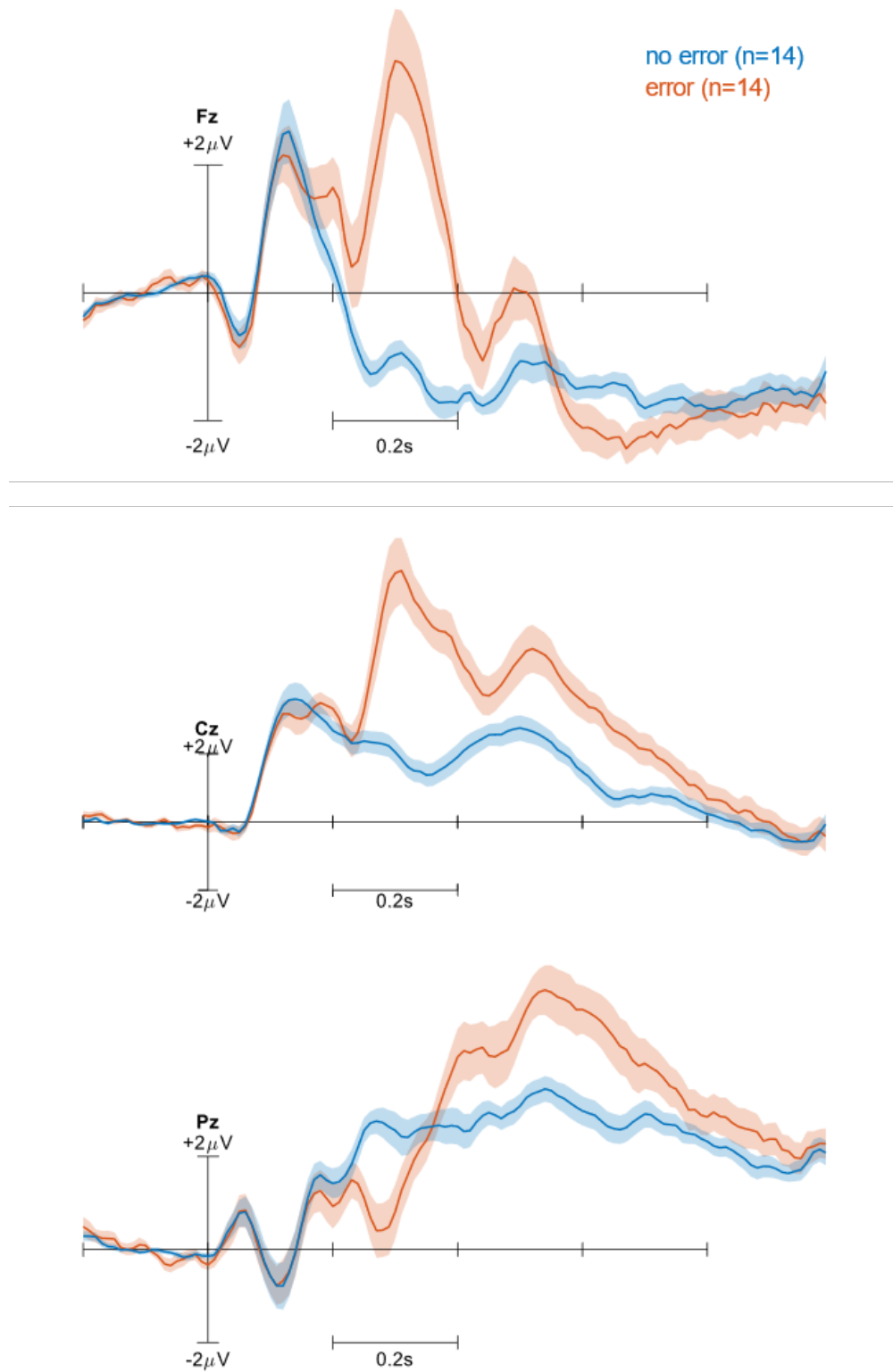


Fig. 3.16 Grand average ERPs ($n=14$) of data cleaned from eye activity through ICA. Curves show error and no error trials with standard deviations from electrode sites Fz, Cz and Pz from the first recording session of the experiment.

Classifier calibration For all 14 participants the obtained classification error rates derived from the data cleaned from eye activity are shown in table 3.5. The average classifier calibration error was 8.02 percent (SD=2.05 percent).

Table 3.5 Classifier calibration error rates in percent with standard deviations of the subject-specific classifiers from recording session 1 (calibration data was cleaned from eye activity)

subject	calibration error	SD
1	8.43	1.15
2	9.03	0.89
3	3.33	2.50
4	5.67	1.56
5	7.29	0.98
6	13.28	4.60
7	9.03	1.27
8	7.54	2.14
9	4.55	0.58
10	9.82	2.26
11	20.37	2.20
12	6.93	3.73
13	5.16	2.33
14	1.86	2.46
mean	8.02	2.05

Classifier visualisation The normalized grand average patterns from each 50 ms time window from 150 to 500 ms of the classifier models of all 14 subjects are shown in figure 3.17. The patterns show that the classifiers have the highest weights in the time window 300-350 ms. Here the focus is dipolar on central and (right lateral) occipital sites. A similar weight distribution over sites is seen in the time window 250-300 ms with lower overall weights. In the time window 150-200 ms the classifiers also put strong weights on occipital sites.

Further analysis with the source localisation tool showed that indeed sources used for the classifiers' decisions now are mainly in a central unilaterized part of the cortex, roughly in the area where the anterior cingulate cortex (ACC) is located. The strongest weights are placed on the fourth time-window from 300 to 350 ms again. Source localisation for all seven 50 ms time windows from 150 to 500 ms are depicted in figures 3.18 to 3.24.

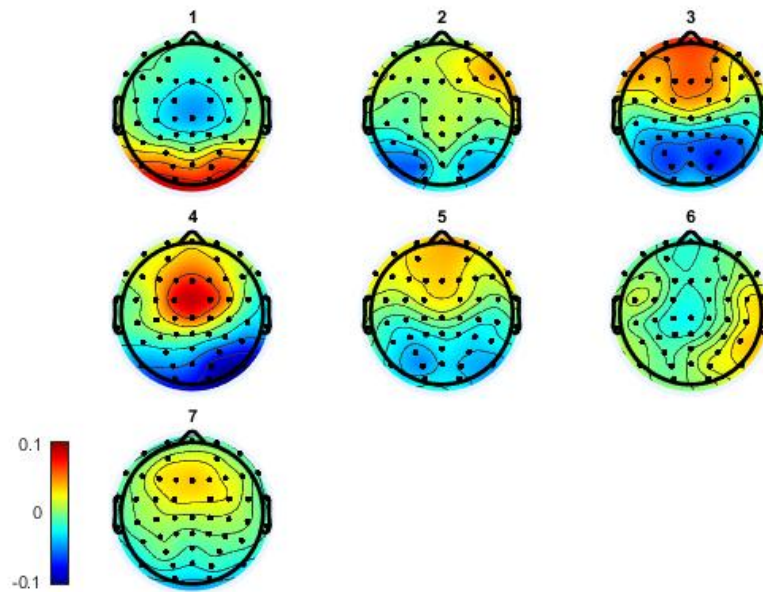


Fig. 3.17 Grand average patterns from each 50 ms long time window from 150 to 500 ms of the classifier models of all 14 subjects trained on data cleaned from eye activity

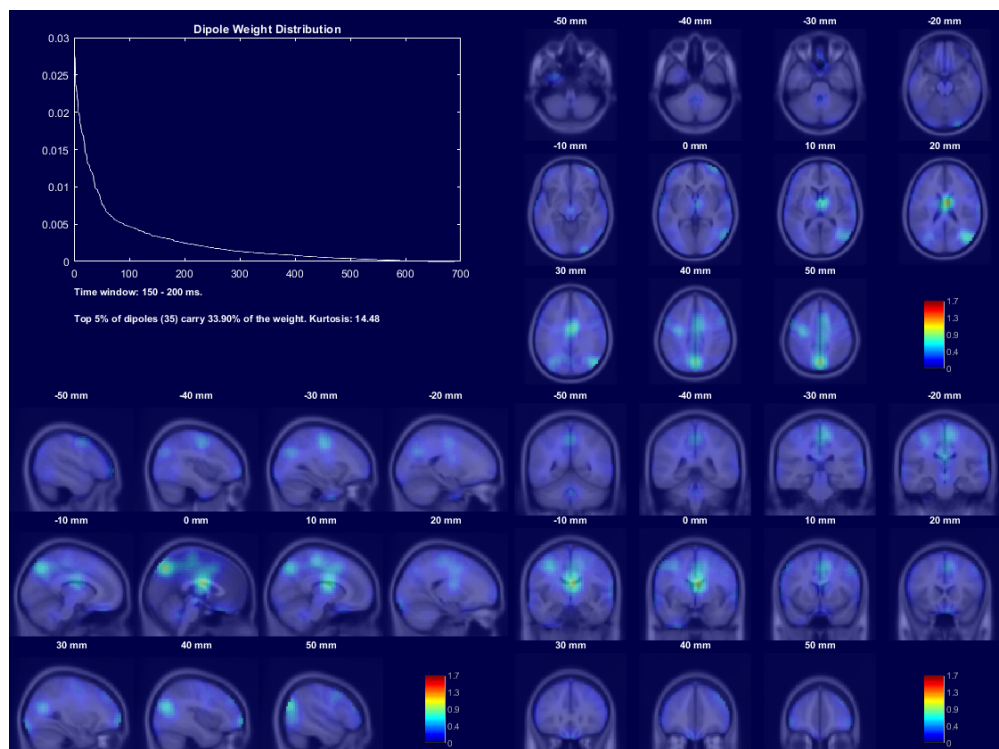


Fig. 3.18 Visualisation of the subject-specific trained on data cleaned from eye component activation in time window 150-200 ms: Weighted dipole density plots showing the relevance of cortical areas to the windowed means classifier. Slices are labelled with their corresponding MNI coordinates. Top left: sorted dipole weight distribution.

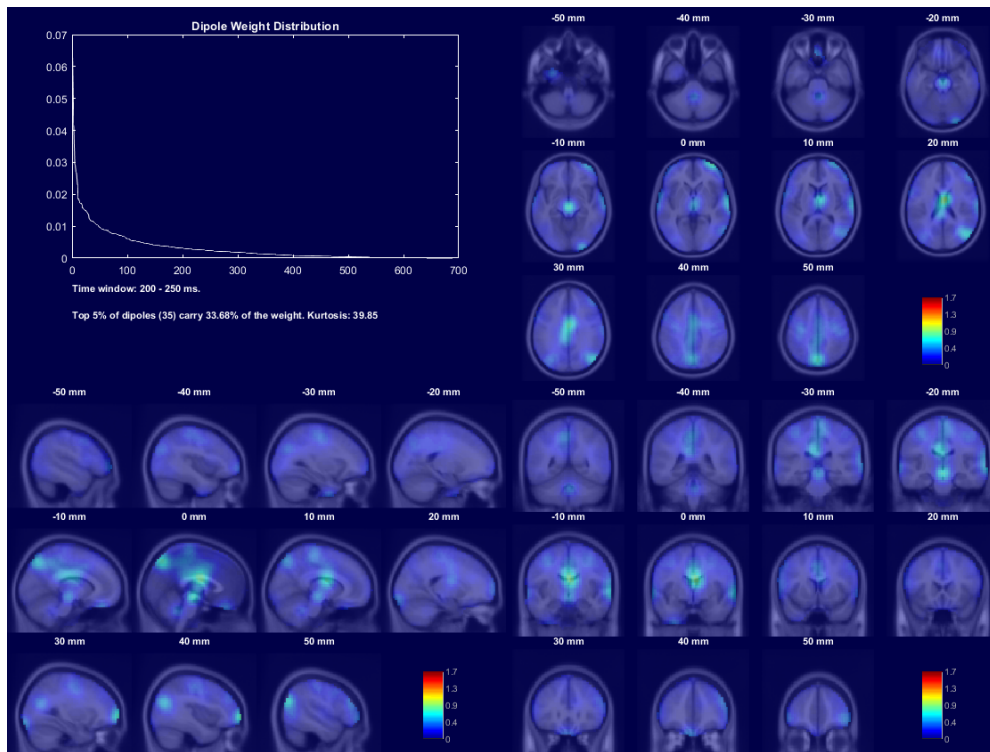


Fig. 3.19 Visualisation of the subject-specific classifier trained on data cleaned from eye component activation in time window 200-250 ms.

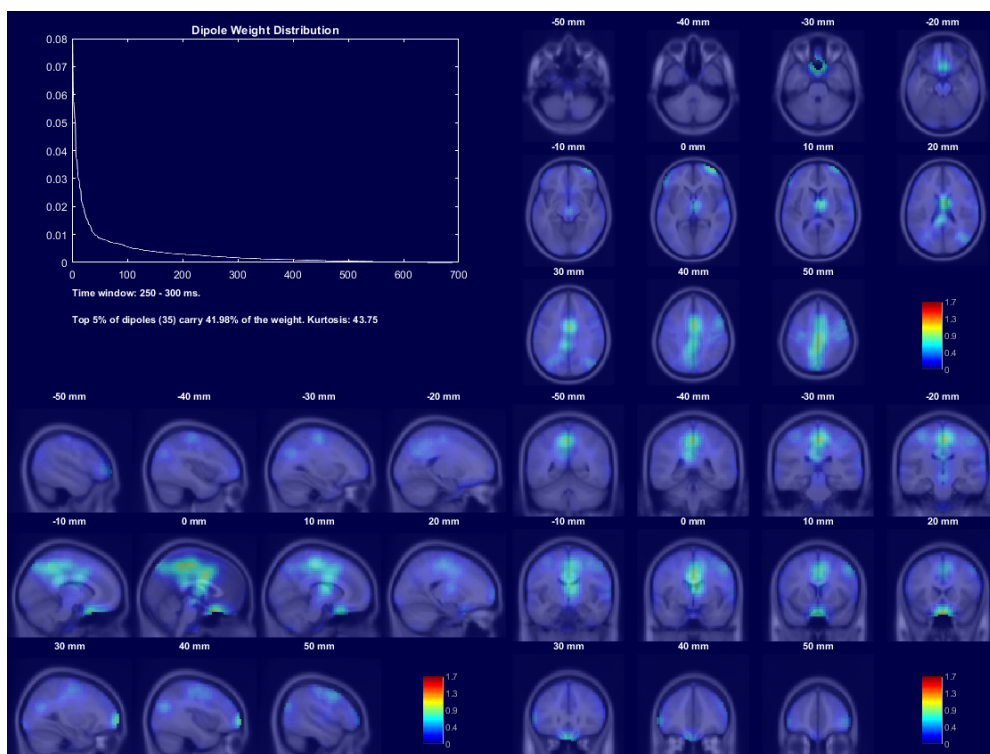


Fig. 3.20 Visualisation of the subject-specific classifier trained on data cleaned from eye component activation in time window 250-300 ms.

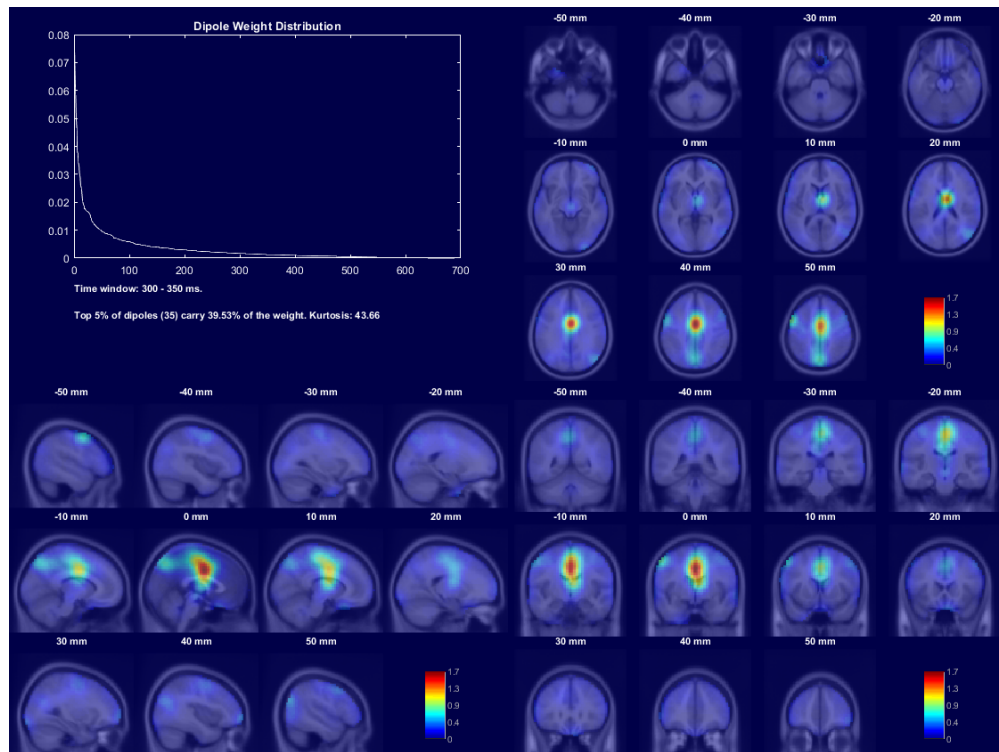


Fig. 3.21 Visualisation of the subject-specific classifier trained on data cleaned from eye component activation in time window 300-350 ms.

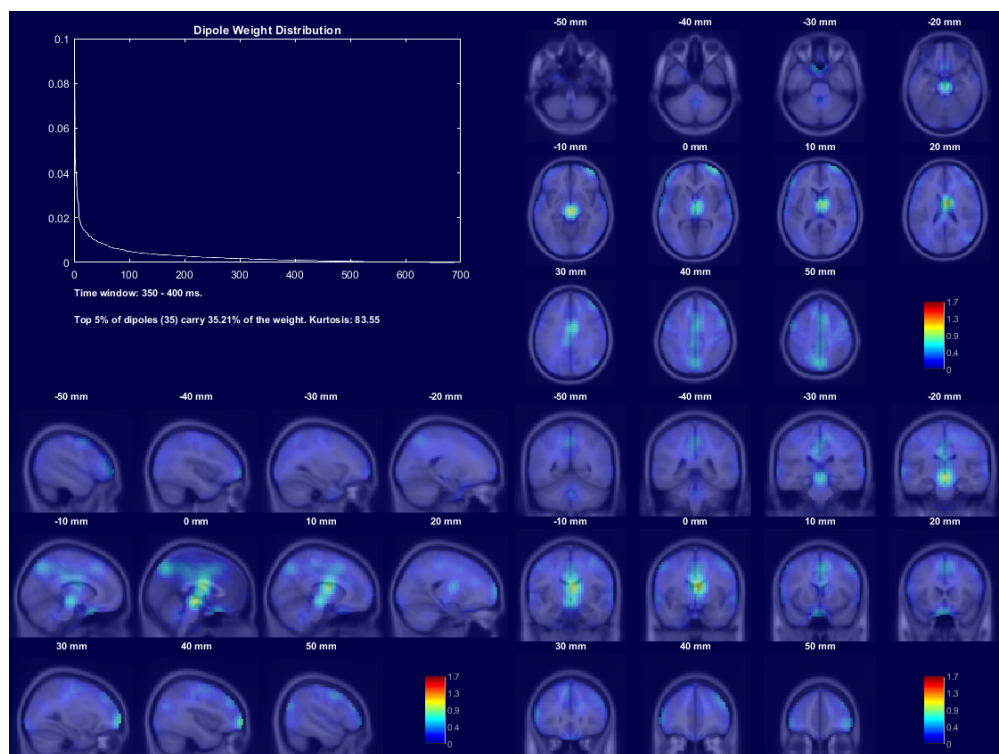


Fig. 3.22 Visualisation of the subject-specific classifier trained on data cleaned from eye component activation in time window 350-400 ms.

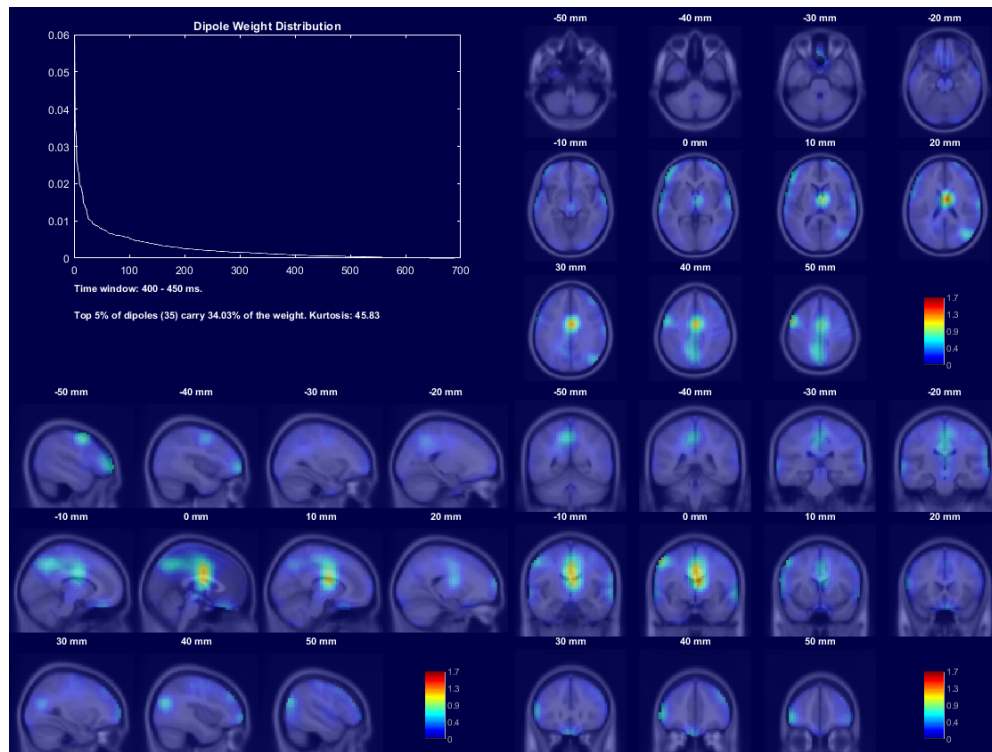


Fig. 3.23 Visualisation of the subject-specific classifier trained on data cleaned from eye component activation in time window 400-450 ms.

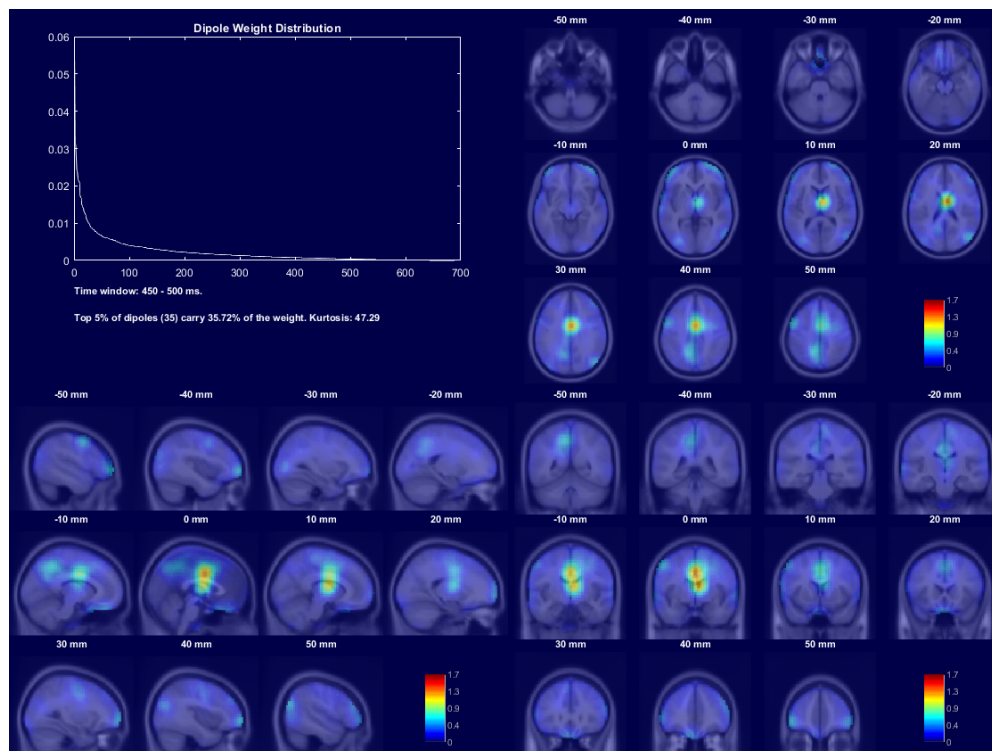


Fig. 3.24 Visualisation of the subject-specific classifier trained on data cleaned from eye component activation in time window 450-500 ms.

Classifier application to testdata We applied the subject-specific classifiers trained on data cleaned from eye component activity to testdata from the fourth recording block from this recording session. Here the average prediction error was at 19.41 percent (SD=9.39 percent). The obtained prediction errors for each subject are shown in table 3.6.

Table 3.6 Classifier prediction errors rates in percent from application of subject-specific classifiers to testdata from recording session 1 (calibration data was cleaned from eye activity)

subject	prediction error
1	12.33
2	13.56
3	18.41
4	10.90
5	43.81
6	21.78
7	22.57
8	11.08
9	29.49
10	23.95
11	25.47
12	13.68
13	15.27
14	9.48
average	19.41

3.2.2.3 Discussion

Overall the analyses in this second part of the study showed that we can train a subject-specific detection classifier, which bases its decisions on cortical activity, instead of mostly on eye activity.

Calibration error rates were low, although not as low as in study part 1. This loss of classifier performance could have different reasons. It could be that eye activity is highly different for erroneous and correct trials and that this information is not available to the classifier anymore, when it is calibrated in data cleaned from eye activity. Another reason might be, that the classifier trained on data without eye activity has trouble making predictions on new (test-) data which is more noisy as it was not cleaned from eye activity.

The application error rate of the new classifier models however is still low at an average of 19 percent, which should be suitable for many HMI applications.

Classifier visualisation showed, that the classifiers calibrated on the data cleaned from eye activity now use cortical sources for classification decisions. These cortical sources seem to be located in the region of the ACC, which has been reported in several neurophysiological studies as a modulator of error perception processes (see section 2.6.2). This validation shows that these new models are truly brain-computer interfaces that work on cortical activity and not on another physiological modality such as e.g. eye activity. To obtain these classifier models, an ICA solution must be computed though which can be time consuming and therefore not suitable for everyday BCI application in real-world contexts.

This part of the study showed that we can train a subject-specific error detection classifier, which bases its decisions on cortical activity. For this classifier, however, it is still required to collect subject-specific training data for classifier calibration in every session when the classifier should be used online by the specific subject. Therefore, in a third part of this study, we investigated, whether the here generated subject-specific classifier models from the first recording session, can be applied to new data of the same person in a new recording session from a different day.

3.2.3 Part 3: Session independence - Applying subject-specific classifier models from a first recording session to new subject-specific testdata from a second recording session

In this third part of the study we applied subject-specific classifier models trained in part 2 of this study to new data from the same subject recorded on a different recording session on a different day. For each subject we compared the performance of the classifier model on the new data to its performance on testdata from the first recording session. The results of these analyses give conclusions to whether the subject-specific error-detection classifier models we defined in part 2 of the study are session-dependent or -independent.

3.2.3.1 Methods

Participants For this second recording session of the study, we recorded new data from the same 14 subjects as in the first recording session.

Apparatus EEG data was again recorded from 50 electrode sites, now with a sampling rate of 100 Hz. EOG was recorded additionally from 4 electrodes. We used Brain Products EasyCap electrodes, with electrode placements according to the 10/20 system. The reference was the nasion and the ground electrode at Fpz. All electrode impedances were kept below 20k Ω . Data was recorded in two blocks: The subject's tasks were the same as in recording blocks three and four of recording session 1 (see 3.2.1.1). Hence, in the first recording block of this second recording session, subjects played the RLR-game against a trained opponent without any machine errors induced. In the second block, the subject again played against the opponent while machine errors were induced into the RLR-game for around 30 percent of trials. In each of the two recording blocks, approximately 40 rounds were played.

Data Analysis Subject-specific classifier models trained in part two of this study - trained on data cleaned from eye component activity - were applied to the second data block recorded in this second recording session to obtain classifier prediction error rates as an estimate of the classifiers' online performance.

The testdata were again pre-processed by first removing EOG channels. We then set a lowpass filter of 200 Hz and rereferenced to the common average.

3.2.3.2 Results

In this section, we present the results of application of the subject-specific classifier models based on training data from the first recording session, to new subject-

specific data from a new second recording session.

The average classifier application error rate when subject-specific classifiers from the first recording session are applied to new data from the second recording session was at 20.43 percent (SD = 11.01 percent). The specific subject's classifier application error rates in percent are shown in table 3.7.

Table 3.7 Classifier prediction error rates in percent from application of the subject-specific classifiers from the first recording session to testdata from recording session 2

subject	prediction error
1	17.95
2	16.52
3	13.23
4	14.95
5	32.07
6	25.11
7	16.02
8	7.78
9	41.98
10	33.98
11	32.40
12	16.22
13	14.45
14	3.37
average	20.43

Subject-specific classifier prediction error rates from application to testdata from recording sessions 1 and 2 were subjected to a paired t-test. The test showed, that prediction error rates were not significantly different on testdata from the first recording session (M=19.41, SD=9.39) than on testdata from the second recording session (M=20.43, SD=11.01), $t(13)=0.55$, $p = 0.591$.

Discussion

In this part of the study, we investigated, if the subject-specific error-detection classifier models could be transferred from one recording session to another. We used the subject-specific classifier models trained on data from recording session 1 and applied them to testdata from recording session 1 and to data from a new recording session 2. In the new recording session data was recorded from the same subjects who performed the same task only on a different day. The results showed no signif-

icant difference between the application of subject-specific classifiers to testdata from the same or a different recording session. The average classifier application error rate in testdata from both recording sessions was about 20 percent. We therefore conclude, that the subject-specific error-detection classifier can be transferred from one recording session to another with similar low prediction error rates.

This *session-independence* can be very beneficial for real-world use of BCIs, as it means that the user, here the specific subject, only needs to calibrate the error-detection classifier in an initial recording session and can then in other sessions omit the classifier calibration phase to start using the BCI for support right away. This would save a lot of the time usually needed for repeated classifier calibration in every session, which is usually a bottleneck when it comes to real-world applicability of BCI systems.

In the next part of our investigation, we will examine whether a session independent classifier can be trained on data from multiple subjects as well. We will train the error-detection classifier not only on data of one specific subject from recording session 1. Instead, we will train a *group classifier* on all calibration data of recording session 1 and then apply it to the subject-specific testdata of recording session 2. We will further investigate whether this group classifier is working on the same cortical sources for classification as the subject-specific classifiers we examined before.

3.2.4 Part 4: Session independence - Training a group classifier on calibration data from the first recording session and applying it to new subject-specific test-data from a second recording session

In this part of the study we examined, if a session independent classifier can be trained on data from not only one, but multiple subjects. We tested if this *group classifier* can be applied to testdata from a specific subject with prediction error rates, that are comparable to a subject-specifically trained classifier model. To this end, we trained the group classifier on all calibration data of recording session 1 and then applied it to the subject-specific testdata of recording session 2. We further investigated if the group classifier works on similar cortical sources for classification as the subject-specific classifiers we examined before.

3.2.4.1 Data Analysis

We used the calibration data of recording session 1 (block 1 and 2) from all 14 subjects to train the group classifier. The pre-processing was the same as in part 2 of this study and we used all 50 EEG channels of the data. The classification approach remained the same as in the parts before with the same parameters.

We then applied this obtained error-detection group classifier model to the subject-specific testdata (block 2) from the second recording session, yielding prediction error rates for every subject. These prediction error rates were subjected to a paired t-test, comparing them to prediction errors obtained from applying the subject-specific classifier models trained on data from recording session 1 to the same testdata from recording session 2 (see study part 3 and table 3.7).

Neurophysiology of the group classifier was analysed as well - we regarded patterns from each 50 ms time window from 150 to 500 ms of the group classifier model. We also examined the weighted dipole density plots for the group classifier showing the relevance of cortical areas to the windowed means classifier.

3.2.4.2 Results

Group classifier calibration The calibration error of the group classifier was at 18.06 percent (SD=3.46 percent).

Group classifier application The average prediction error for application of the group classifier to subject-specific testdata from recording session 2 was at 22.23 percent (SD=6.66 percent). The subject-specific prediction error rates are shown in table 3.8. Results of the paired t-test showed that prediction error rates of the group classifier applied to testdata from recording session 2 (M=22.23, SD=6.66)

were not significantly different from prediction error rates obtained from subject-specific classifier models ($M=20.43$, $SD=11.01$), $t(13)=1.22$, $p = 0.245$.

Table 3.8 Classifier prediction error rates in percent from application of the group classifier from the first recording session to subject-specific testdata from recording session 2

subject	prediction error
1	18.97
2	15.04
3	17.46
4	20.10
5	31.52
6	29.22
7	20.39
8	17.22
9	34.57
10	25.10
11	29.05
12	20.00
13	20.81
14	11.80
average	22.23

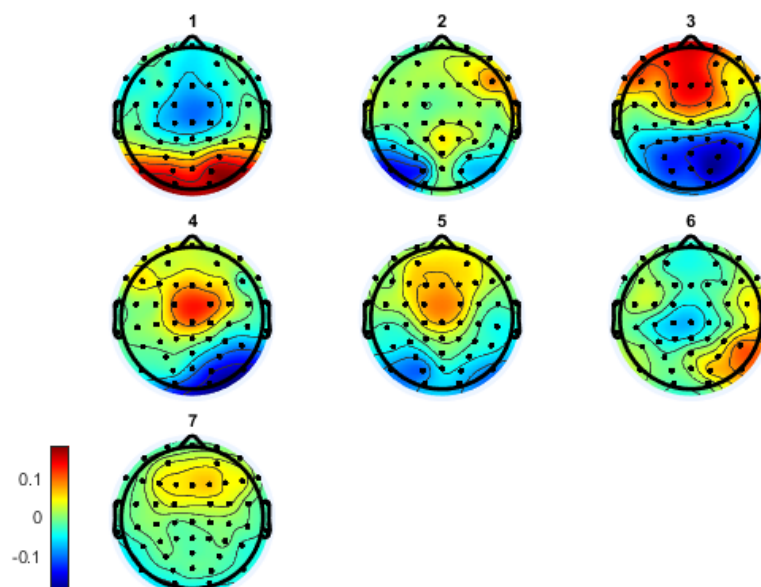


Fig. 3.25 Patterns from each 50 ms long time window from 150 to 500 ms of the group classifier model trained on data cleaned from eye activity.

Group classifier visualisation The patterns from each 50 ms time window from 150 to 500 ms of the group classifier model are shown in figure 3.25. Similar to the grand average pattern of the subject-specific classifiers, the group classifier model shows strong weights on central and (right lateral) occipital electrode sites in the time window 300 to 350 ms, and additionally in the time window from 250 to 300 ms. A topographically similar pattern is seen in the first time window from 150 to 200 ms, with strong occipital weights but less pronounced weights on central sites.

Further analysis of the group classifier model with the source localisation tool shows that the group classifier focuses on similar cortical areas close to the ACC, with the highest weights in time window four from 300 to 350 ms again. Weighted dipole density plots showing the relevance of cortical areas to the group classifier are shown in figures 3.26 to 3.32.

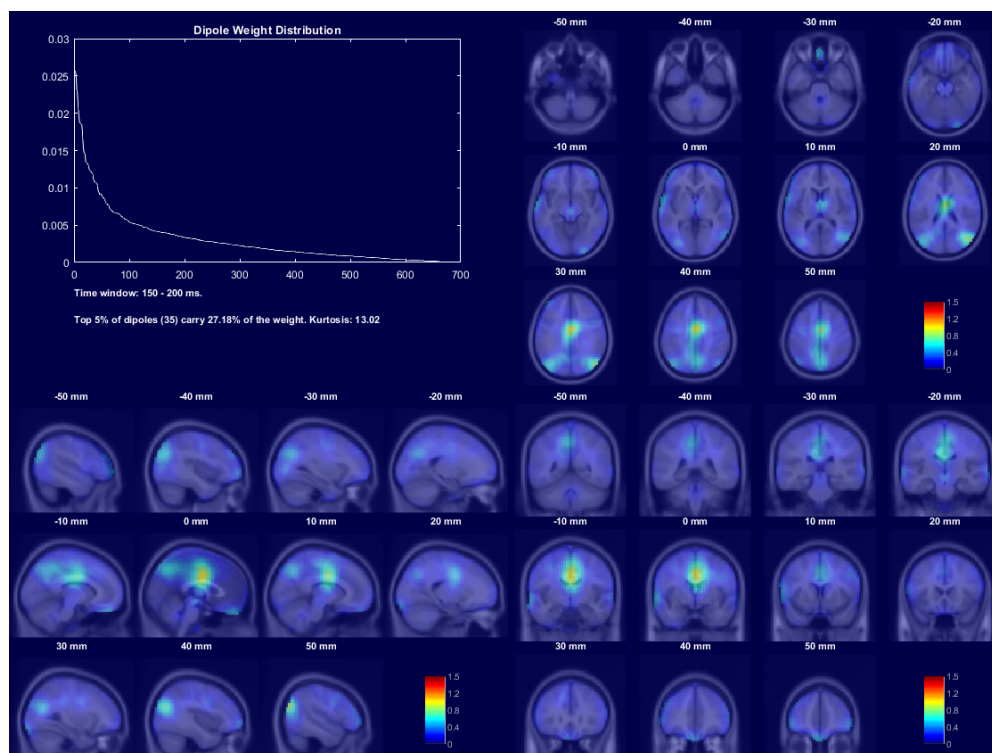


Fig. 3.26 Visualisation of the group classifier trained on data cleaned from eye component activation in time window 150-200 ms: Weighted dipole density plots showing the relevance of cortical areas to the windowed means classifier. Slices are labelled with their corresponding MNI coordinates. Top left: sorted dipole weight distribution.

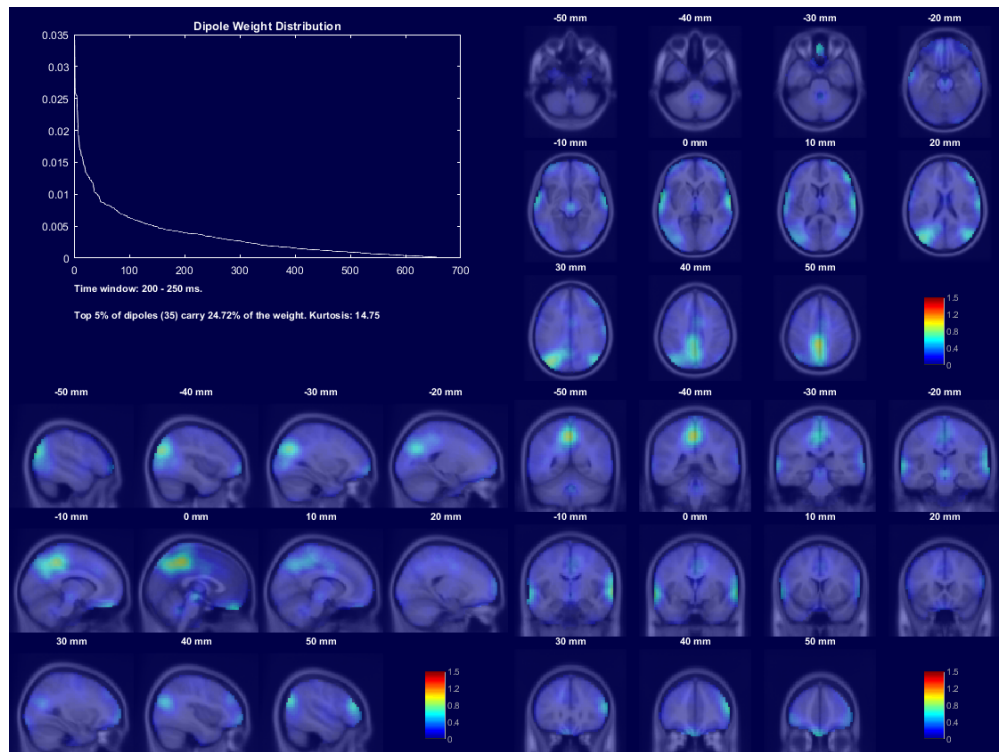


Fig. 3.27 Visualisation of the group classifier trained on data cleaned from eye component activation in time window 200-250 ms.

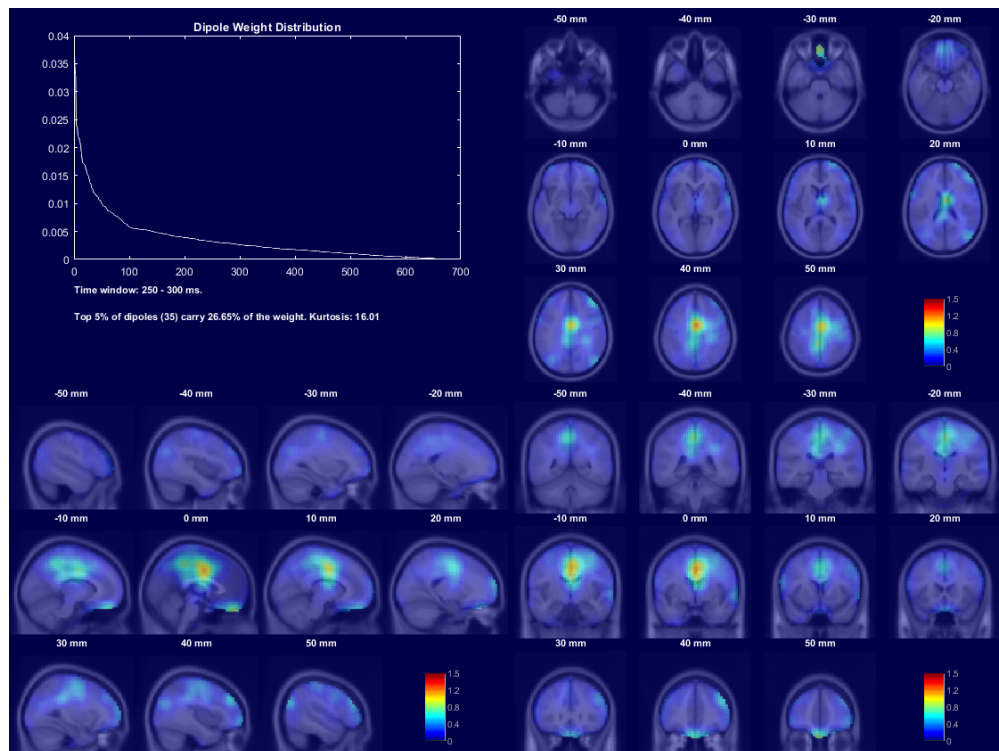


Fig. 3.28 Visualisation of the group classifier in time window 250-300 ms.

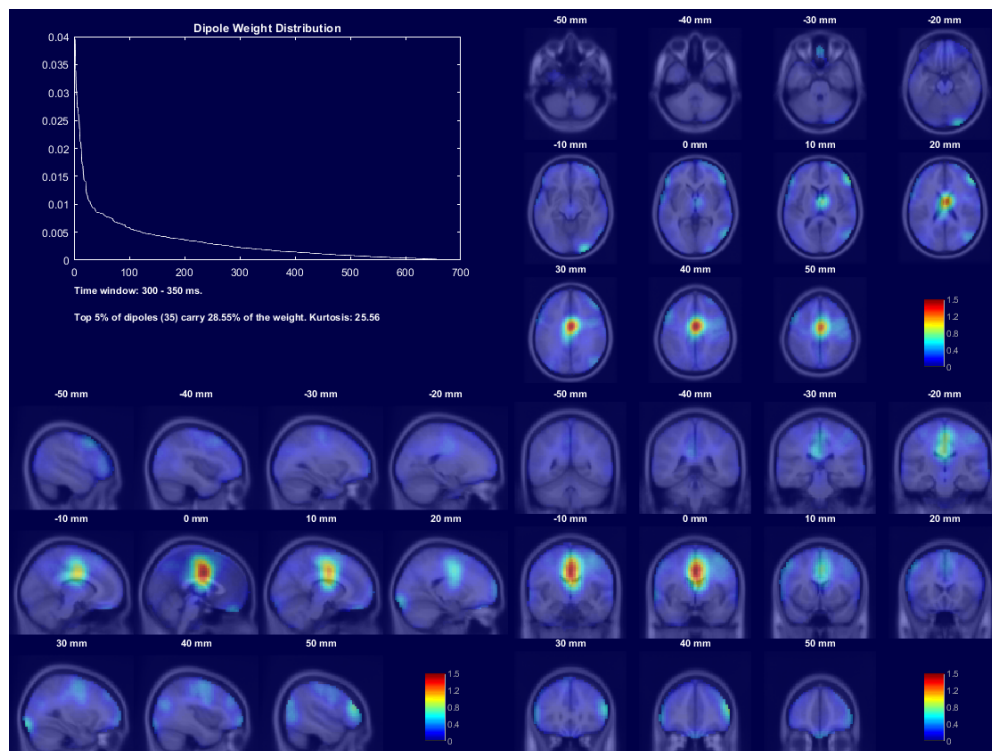


Fig. 3.29 Visualisation of the group classifier in time window 300-350 ms.

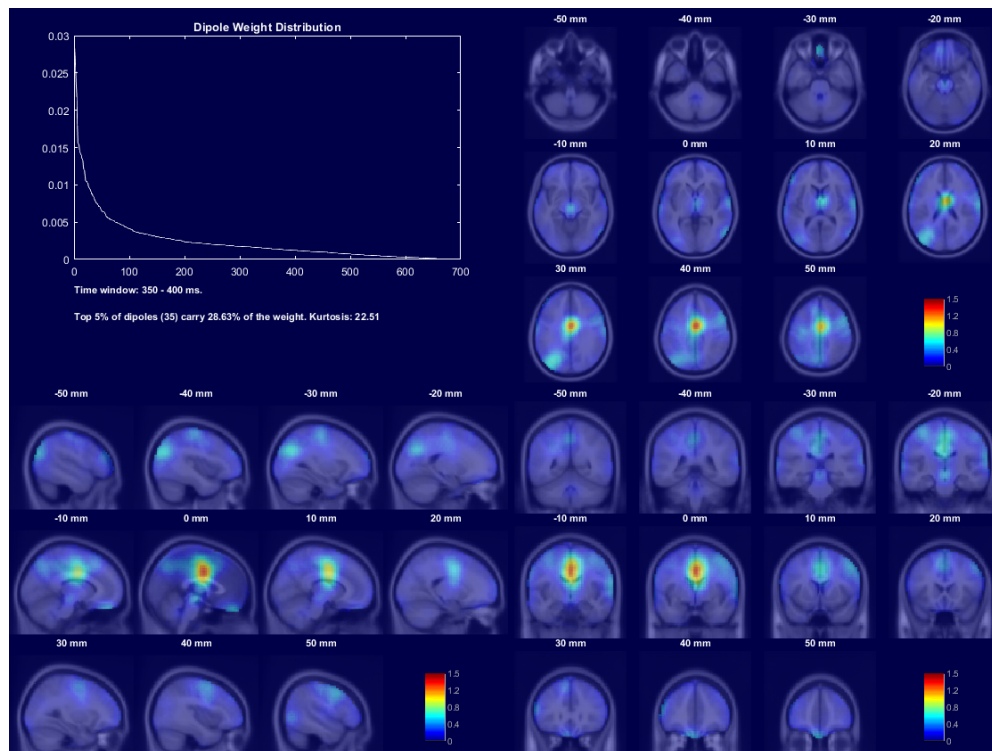


Fig. 3.30 Visualisation of the group classifier in time window 350-400 ms.

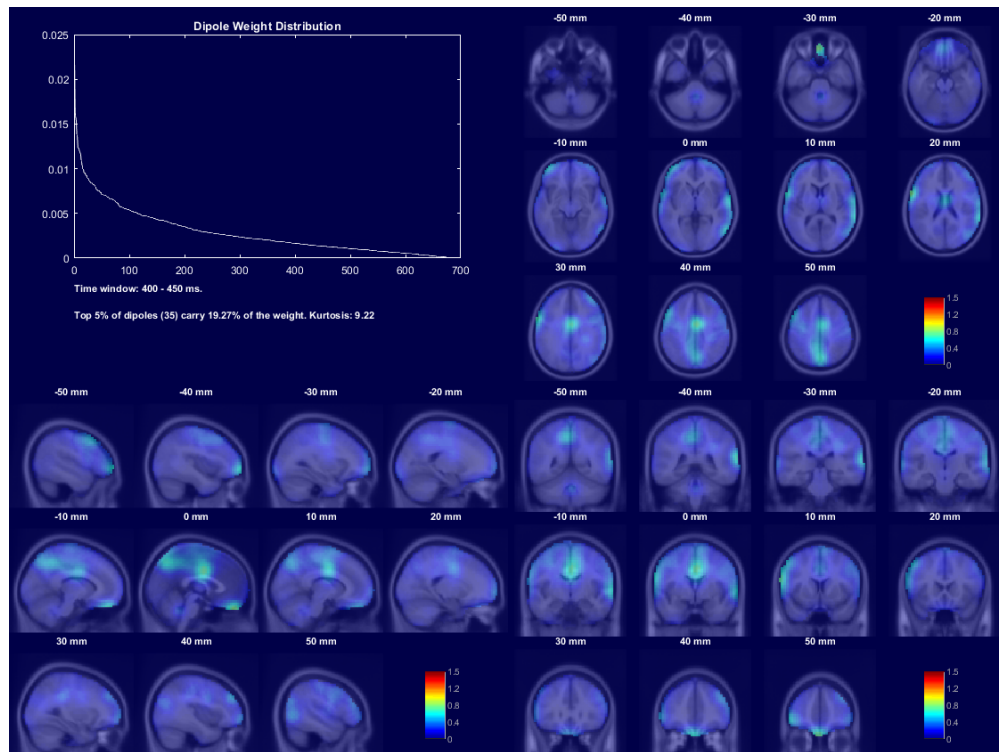


Fig. 3.31 Visualisation of the group classifier in time window 400-450 ms.

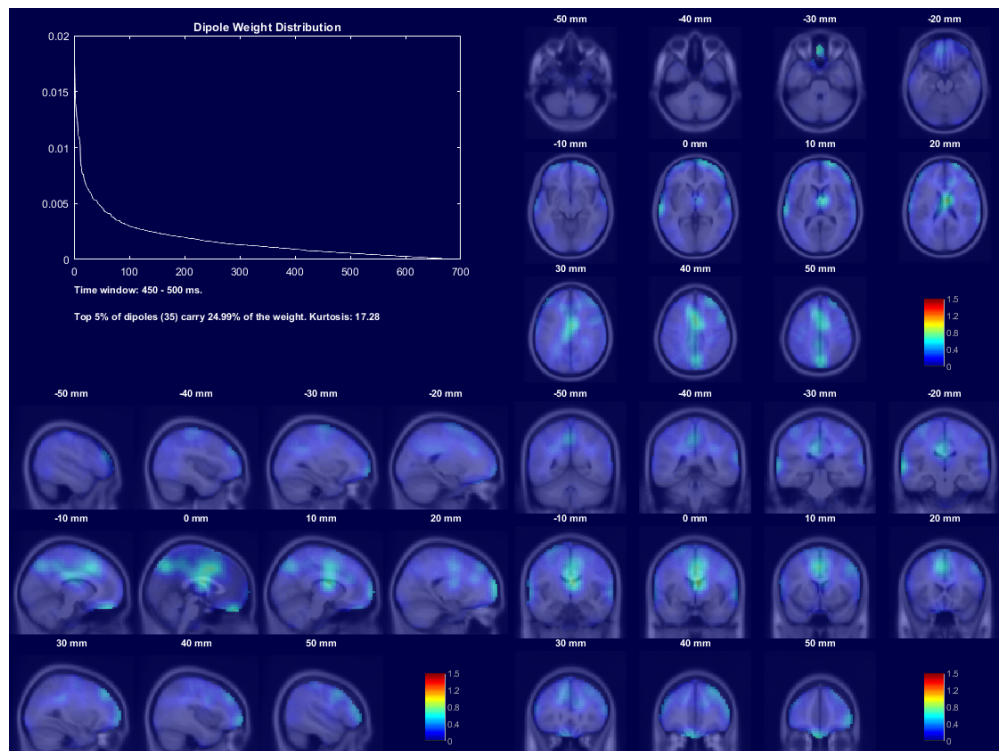


Fig. 3.32 Visualisation of the group classifier in time window 450-500 ms.

3.2.4.3 Discussion

In this part of the study, we trained a group classifier on all calibration data from the first recording session. We applied this classifier to new subject-specific test-data from a different recording session with the same subjects, then we compared its performance to that of the subject-specific classifier models. Results showed no significant difference between the prediction error rates of the group classifier and subject-specific classifiers. We also compared classifier model visualisations between the group and subject-specific classifiers and found strong similarities in classifier patterns, as well as in the weighted dipole density plots showing the relevance of cortical areas to the windowed means classifier. In both cases, the classifier weights are strongest in the time window from 300 to 350 ms, which is where the error positivity (Pe) occurs. The cortical sources underlying the classifier's decision in this time window are around the area of the ACC which is known to be involved in error-processing.

These results suggest, that instead of training a classifier model, which is based on data from only a single subject, we can alternatively train one classifier model on data from a group of subjects and yield similar prediction error rates as with subject-specifically trained models. Across subjects, this session-independent group classifier seems to work on the same discriminative features as the subject-specifically trained classifier models. This could indicate, that neural responses to error perception are highly similar across subjects.

The here trained group classifier however is partially informed about data from the specific subject whose new testdata we applied it to, as the group classifier is trained on calibration data from all subjects - including the test subject.

Going one step further towards investigating not only the here demonstrated session-independence of the error-detection classifier, but also *subject-independence*, in the next part of the study, we will apply subject-specific classifier models to test-data from other subjects. We will further train group classifiers, excluding a specific subject's calibration data from the group classifier calibration, and then apply this *leave-one-subject-out* group classifier to the testdata of the subject, who was left out in the calibration dataset.

3.2.5 Part 5: Subject-independence - Applying subject-specific and group classifier models to subject-specific testdata from a different subject

In this part of the study we investigate *subject-independence* as a feature of the subject-specific and the group classifiers we have trained. We apply both classifier types to data from other subjects, whose data was not used for training the classifier's model in the calibration phase.

3.2.5.1 Methods

We firstly applied the subject-specific classifiers trained on calibration data from recording session 1 to testdata of all other subjects of recording session 1. This yields 14 subject-specific classifiers \times 13 test subjects = 182 prediction error rates.

Further, we calibrated 14 new group classifiers on the calibration data from recording session 1. Using the *leave-one-subject-out* scheme, for each of the 14 subjects, their subject-specific calibration data was left out of the calibration and the model was trained on the calibration data of the remaining 13 subjects. The obtained group classifiers were then applied to the testdata of the one specific subject, whose calibration data was not part of the group classifier's training data.

3.2.5.2 Results

Application of subject-specific classifiers to testdata from other subjects The average application error rate of application of the subject-specific classifiers trained on calibration data from recording session 1 to testdata of all other subjects of recording session 1 was at 33.14 percent (SD=5.34 percent). All subject-specific 183 prediction error rates for the specific subjects are shown in table 3.10. We subjected the average application error rates per subject-specific classifier model to a paired t-test, comparing them to the application performance of the subject-specific classifier performance on the same subject's testdata from recording session 1 (see part 2 table 3.6). Results showed, that prediction error rates for subject-specific classifier application to other subjects' testdata (M=33.14 percent, SD=5.34 percent) was significantly worse than prediction error rates of the subject-specific classifier performance on the same subject's testdata (M=19.41 percent, SD=9.39 percent), $t(13)=-4.76$, $p < 0.001$.

Leave-one-subject-out group classifier calibration and application The average calibration error rate of the leave-one-subject-out group classifier models was at 17.97 percent (SD=3.74 percent). The calibration error rates for each subject are shown in table 3.9.

Table 3.10 Classifier prediction error rates in percent from application of the subject-specific classifiers from the first recording session to all other subjects' testdata from recording session 1

classifier subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14
subject	prediction error rates													
1	/	52.67	25.00	27.67	32.67	32.33	37.67	34.00	34.00	30.00	37.33	30.67	27.00	34.00
2	30.28	/	18.30	24.92	29.34	29.97	38.49	28.08	26.81	25.24	28.71	27.13	30.60	27.44
3	33.21	41.88	/	21.30	42.60	23.10	30.69	32.49	36.82	33.21	44.77	20.22	27.08	33.94
4	48.72	39.74	25.64	/	24.36	39.74	30.77	26.28	23.08	25.00	42.95	23.08	25.64	17.95
5	48.25	53.02	50.16	23.81	/	26.03	32.38	48.57	37.46	40.95	56.19	24.44	26.35	36.19
6	43.79	42.86	38.41	40.75	42.62	/	44.50	38.41	40.52	37.94	48.24	39.81	37.70	34.19
7	29.78	34.80	29.78	25.71	33.23	26.96	/	27.59	35.74	30.41	32.92	26.02	31.03	28.21
8	33.54	48.42	28.80	18.35	28.16	31.01	31.96	/	28.48	30.06	31.01	28.80	29.43	29.11
9	51.19	43.73	33.22	21.02	51.19	21.69	31.53	48.14	/	48.47	49.49	26.10	25.08	33.56
10	37.72	39.22	25.15	25.15	33.83	28.74	33.23	31.44	24.25	/	29.94	22.46	26.05	20.36
11	44.57	44.57	36.33	30.34	31.46	44.57	40.07	35.21	37.45	27.34	/	24.34	29.59	30.34
12	45.61	50.88	36.49	23.16	38.95	28.42	30.88	37.54	43.16	38.25	46.67	/	22.46	34.04
13	37.40	45.42	23.66	24.81	46.95	32.82	23.28	37.02	39.69	33.21	43.89	23.28	/	28.63
14	40.28	30.33	25.59	18.48	29.38	41.71	20.85	25.12	33.18	27.96	17.06	19.91	30.33	/
average	40.34	43.66	30.50	25.04	35.75	31.32	32.79	34.61	33.90	32.93	39.17	25.87	28.33	29.84
grand average	33.14													

Table 3.9 Classifier calibration error rates in percent with standard deviations of the leave-one-subject-out group classifiers from recording session 1

subject	calibration error	SD
1	16.75	4.32
2	16.86	4.86
3	18.18	4.35
4	17.34	4.09
5	18.47	4.77
6	16.19	3.84
7	18.25	4.36
8	18.45	3.92
9	18.73	3.91
10	18.56	3.43
11	17.05	3.92
12	18.62	2.20
13	19.28	2.23
14	18.84	2.17
average	17.97	3.74

We finally applied the leave-one-subject-out group classifiers to test data from recording session 1 of the subject left out in the respective classifier calibration and yielded an average prediction error rate of 26.92 percent (SD=5.72 percent). The specific subjects' prediction error rates are shown in table 3.11. Prediction error rates of these leave-one-subject-out group classifiers were subjected to a paired t-test, comparing them to the prediction performance of the subject-specific classifier on the same subject's testdata from recording session 1 (see part 2 table 3.6). Results showed, that the prediction error rates for leave-one-subject-out group classifiers (M=26.92 percent, SD=5.72 percent) was significantly worse than prediction error rates of the subject-specific classifier performance on the same subject's testdata (M=19.41 percent, SD=9.39 percent), $t(13)=-3.60$, $p = 0.003$.

We also subjected the average application error rates per subject-specific classifier applied to other subjects' testdata to an independent t-test, comparing it to the prediction error rates for leave-one-subject-out group classifiers. Results showed, that the prediction error rates for leave-one-subject-out group classifiers (M=26.92 percent, SD=5.72 percent) were significantly better than prediction error rates of the subject-specific classifier performance on other subjects' testdata (M=33.14 percent, SD=5.34 percent), $t(13)=-2.98$, $p = 0.006$.

Table 3.11 Classifier prediction error rates in percent of applying the leave-one-subject-out group classifiers from recording session 1 to subject-specific test-data of recording session 1

subject	prediction error
1	25.33
2	22.08
3	23.10
4	31.41
5	37.14
6	35.13
7	22.57
8	23.42
9	29.15
10	23.05
11	29.96
12	31.93
13	25.95
14	16.59
average	26.92

3.2.5.3 Discussion

In this part of the study we investigated the classifier models regarding subject-independence. We applied subject-specific and group classifier models to data from other subjects, whose data was not used for training the classifier's model in the calibration phase. The analyses showed, that both, the subject-specific and group classifier models perform worse on new data from a different subject, than the subject-specific classifier model, calibrated on data from the same subject. Here on average the subject-specific classifier models performed 13.73 percent worse. Leave-one-subject-out group classifiers models performed on average 7.51 percent worse. This decrease in the classifiers' prediction performance however, should be considered from the aspect, that in practice, both the leave-one-subject-out group classifiers and the subject-specific classifiers that are applied to data from new subjects, would not require *any* calibration time in any recording session from the end-user of the BCI system, as the error-detection classifier model is trained beforehand on data samples from other individuals. There is however also a significant difference between the predictive performance of the leave-one-subject-out group classifiers and the subject-specific classifiers. The prediction error rates on data from a new subject are on average 6.22 percent better when leave-one-subject-out group

classifiers are applied.

We conclude that we can train a *subject-independent* classifier for error-detection, on the basis of collected calibration data from other specific subjects. Here we showed, that training a group classifier on data from multiple subjects yields better predictive classifier performance when data from a new subject is given to the classifier, than when the classifier model is calibrated on data from only one person. The classifier application performance of the subject-independent group classifier decreases though, in comparison to a subject-specific classifier model calibrated on data from the same subject that the model later is applied to. This performance decrease of around 7.5 percent however, might still be a valuable and preferable choice in real-world BCI applications, as the proposed group classifier fully supersedes the calibration phase.

In this part of the study we saw that subject transferability of the error-detection classifier is possible and a subject-independent group classifier can be trained yielding acceptable classifier application performance. In the next two parts of the study we investigated the combination of session- and subject-independence of the error-detection group classifier by transferring it to new data from a third recording session with new subjects.

3.2.6 Part 6: Reducing the number of EEG channels used and omitting channels recording eye activity

In this part of the study we tested, if the number of electrodes of our setup can be reduced, without decreasing classifier performance of the error-detection group classifier. Reducing the number of needed electrodes is crucial for real-world application of the error-detection classifier, as end-users want to spend zero or minimal time setting up and preparing the hardware for neurophysiological measurement through e.g. EEG before they can start using a BCI application to support them in their tasks.

Methods

We decreased the number of our initial 50 EEG electrodes setup in two steps.

First, we reduced the electrodes down to 32 by symmetrically removing electrodes that were positioned between two other electrodes, most likely recording mainly redundant EEG activity.

In a second step, we additionally removed frontal electrodes, which mostly record eye activity. This resulted in a final setup of 24 electrodes. We removed these electrodes, as the group classifier is trained on data which was cleaned from eye activity and hence the classifier is not focusing on eye activity for classification of perceived errors.

The three resulting electrode setups are depicted in figure 3.33.

Data Analysis

For each of the three electrode setups, with 50, 32 and 24 electrodes, we trained a subject-specific classifier for all 14 subjects and one group classifier. The calibration data consisted of all datasets from the first two data recording sessions - data block 1, 2, 3 and 4 for recording session 1 and data block 1 and 2 from recording session 2. All data was cleaned from eye activity by subtracting eye component data identified through ICA. The calibration approach stayed the same as in the other parts of this study (see section 3.2.1.2). We plotted normalized grand average patterns for the subject-specific classifier models and patterns for the group classifier model for each electrode setup. Finally, we subjected subject-specific classifier calibration error rates to a one-way repeated measures ANOVA with a significance level of 0.05, to investigate the effect of the three electrode setups regarding classifier accuracies.

Results

3.2.6.1 Electrode reduction in the recording setup

An overview of the original electrode setup with 50 channels is shown in figure 3.33. The reduced electrode setups with 32 (after symmetrical reduction of the whole setup) and 24 electrodes (after removing eight additional frontal electrodes) are also depicted.

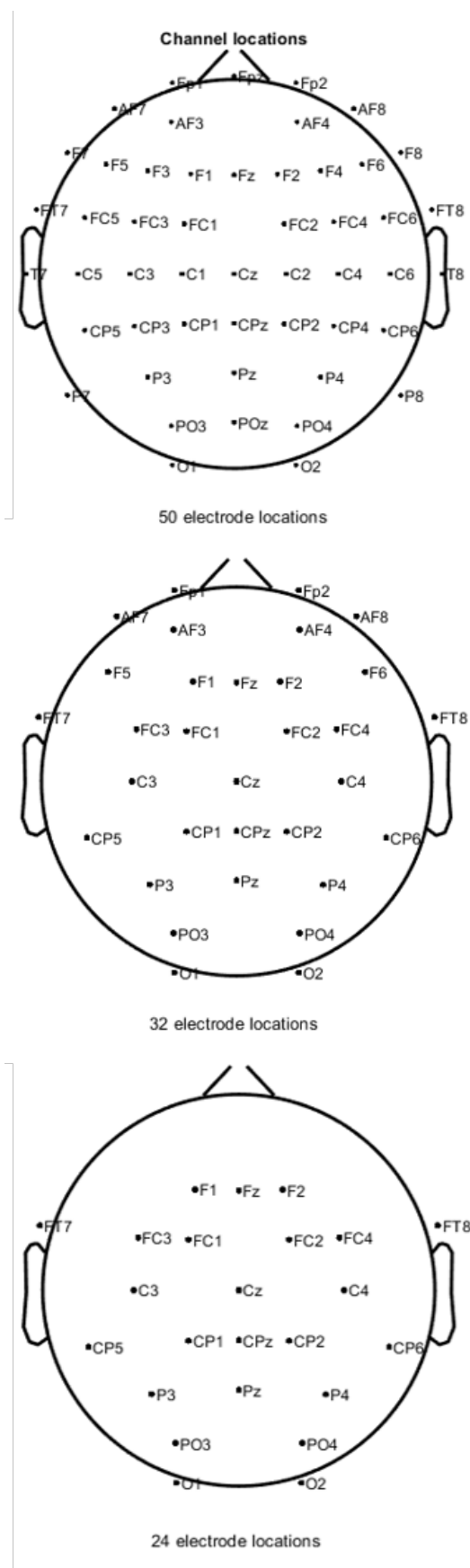


Fig. 3.33 The three different electrode setups: all 50 electrodes from recording sessions 1 and 2, the symmetrically reduced electrode setup with 32 electrodes and the 24 electrode setup with the final omission of frontal electrodes.

Subject-specific classifier calibration with different electrode setups Subject-specific classifier calibration error rates are shown in table 3.12 for all three electrode setups. The respective normalized grand average patterns are depicted in figures 3.34 to 3.36. On average, calibration error rates for the subject-specific classifiers calibrated on 50 electrodes was at 7.27 percent with an average standard deviation of 1.39 percent. For the 32 electrode setup the average error rate was at 8.34 percent (average SD was 1.45 percent) and for the 24 electrode setup the average error rate was at 8.37 percent (average SD was 1.24 percent).

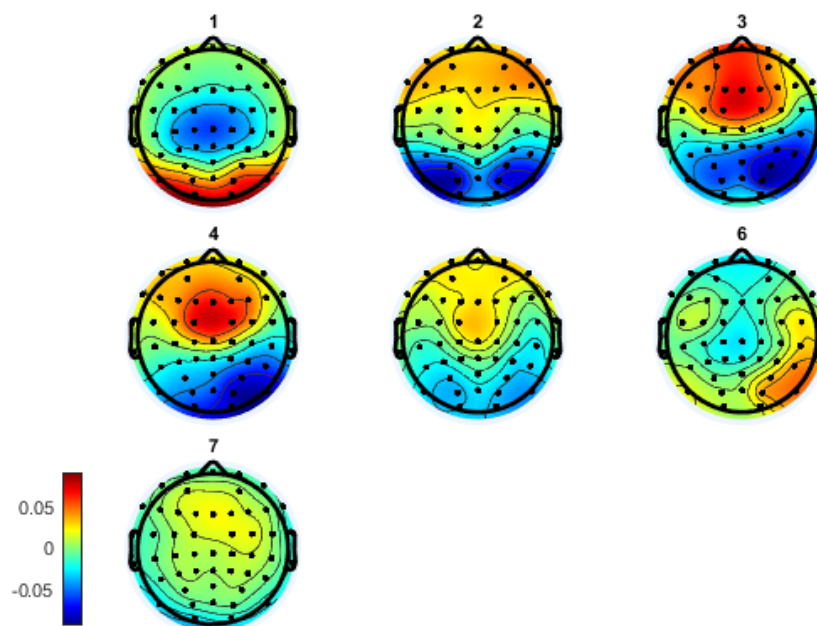


Fig. 3.34 Grand average patterns from each 50 ms long time window from 150 to 500 ms of the classifier models of all 14 subjects trained on data cleaned from eye activity. The calibration data was all data from recording session 1 and 2 and the classifier was trained on the 50 electrode setup.

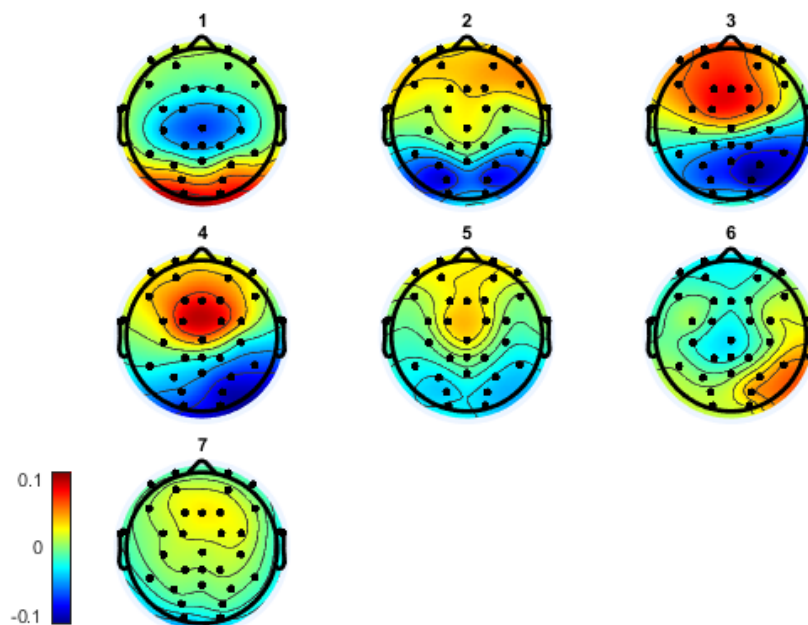


Fig. 3.35 Grand average patterns from each 50 ms long time window from 150 to 500 ms of the classifier models of all 14 subjects trained on data cleaned from eye activity. The calibration data was all data from recording session 1 and 2 and the classifier was trained on the 32 electrode setup.

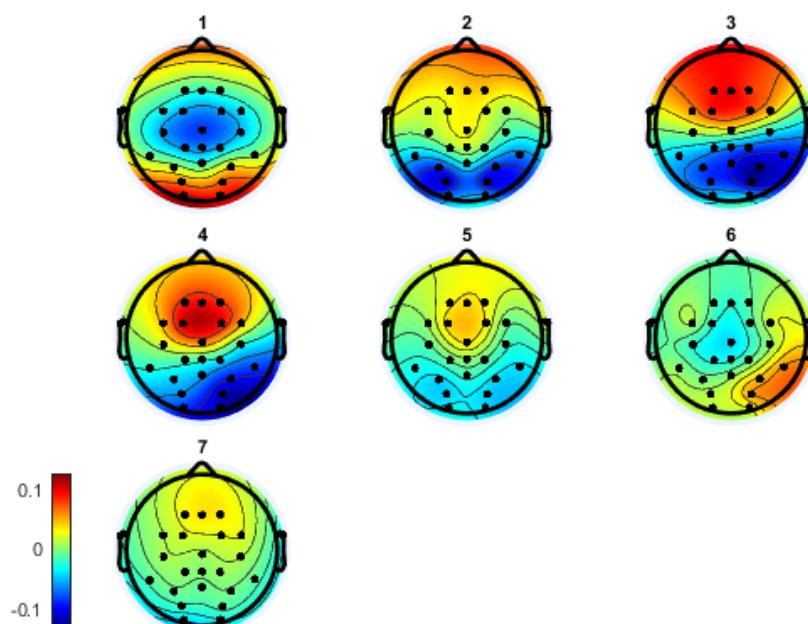


Fig. 3.36 Grand average patterns from each 50 ms long time window from 150 to 500 ms of the classifier models of all 14 subjects trained on data cleaned from eye activity. The calibration data was all data from recording session 1 and 2 and the classifier was trained on the 24 electrode setup.

The repeated measures ANOVA performed on the subject-specific classifier calibration error rates within the three different electrode setups showed, that the electrode setups had a significant effect on the resulting calibration error rates, $F(2, 26) = 21.51$, $p < .001$. Post-hoc paired t-tests with bonferroni corrected alpha levels ($\alpha = 0.017$) were conducted between the data of the three electrode setups. These showed that calibration error rates from the 32 electrode setup were significantly higher than for the 50 electrode setup ($t(13)=5.532$, $p < 0.001$). Calibration error rates from the 24 electrode setup were also significantly higher than for the 50 electrode setup ($t(13)=4.621$, $p < 0.001$). There was no significant difference found between calibration error rates of the electrode setup with 32 and the one with 24 electrodes, $t(13)=0.202$, $p=0.843$.

Table 3.12 Subject-specific classifier calibration error rates in percent and standard deviations of training with a 50, 32 and a 24 electrode setup on all data from recording sessions 1 and 2

subject	50 electrode setup		32 electrode setup		24 electrode setup	
	calib. error	SD	calib. error	SD	calib. error	SD
1	7.75	1.55	9.16	1.21	8.56	0.96
2	7.96	1.53	9.39	2.33	10.11	1.61
3	5.18	1.36	6.94	0.84	7.29	1.11
4	5.41	1.09	5.75	0.93	5.83	0.70
5	5.18	0.80	5.58	1.57	5.50	1.42
6	12.60	1.86	14.77	1.88	15.33	1.60
7	9.68	1.31	10.10	1.43	10.65	1.14
8	5.87	1.44	6.41	0.63	6.33	1.10
9	3.86	0.73	4.18	1.32	4.02	0.91
10	6.30	0.52	8.74	1.86	8.12	0.25
11	15.39	2.16	16.71	1.31	16.33	1.53
12	9.18	2.39	9.93	2.96	9.75	2.92
13	5.40	1.23	6.73	0.91	7.35	1.39
14	2.09	1.46	2.43	1.14	2.00	0.79
average	7.27	1.39	8.34	1.45	8.37	1.24

Group classifier calibration with different electrode setups Classifier calibration accuracies for the group classifiers trained on the three different electrode setups are shown in table 3.13. The respective patterns of these group classifiers are depicted in figures 3.37 to 3.39.

Table 3.13 Group classifier calibration error rates in percent and standard deviations of training with a 50, 32 and a 24 electrode setup on all data from recording sessions 1 and 2

50 electrode setup		32 electrode setup		24 electrode setup	
calibration error	SD	calibration error	SD	calibration error	SD
17,36	2,76	17,15	2,55	17,28	2,90

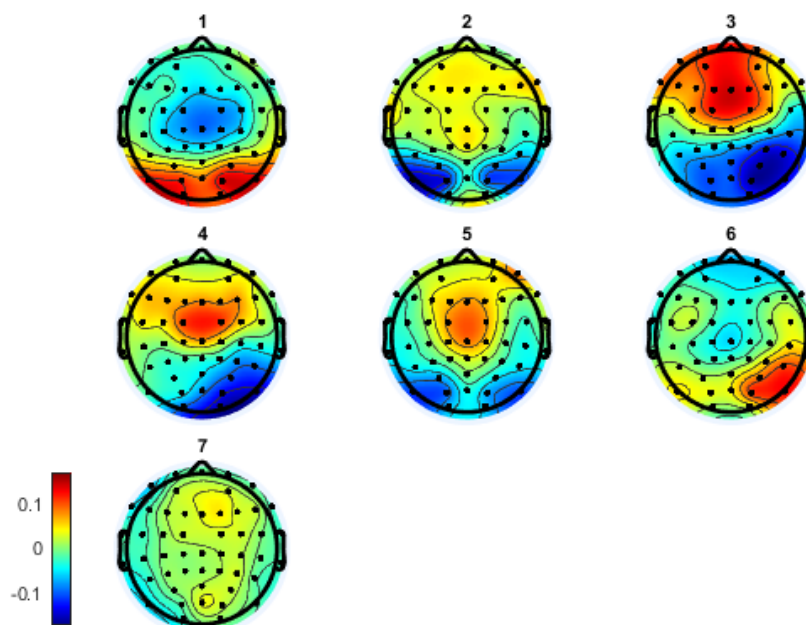


Fig. 3.37 Grand average patterns from each 50 ms long time window from 150 to 500 ms of the group classifier trained on data cleaned from eye activity. The calibration data was all data from recording session 1 and 2 and the classifier was trained on the 50 electrode setup.

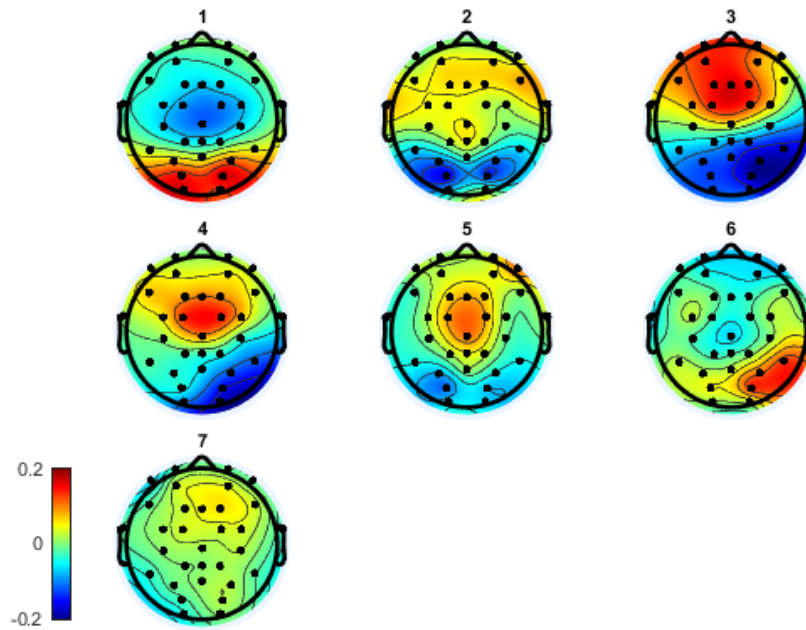


Fig. 3.38 Grand average patterns from each 50 ms long time window from 150 to 500 ms of the group classifier trained on data cleaned from eye activity. The calibration data was all data from recording session 1 and 2 and the classifier was trained on the 32 electrode setup.

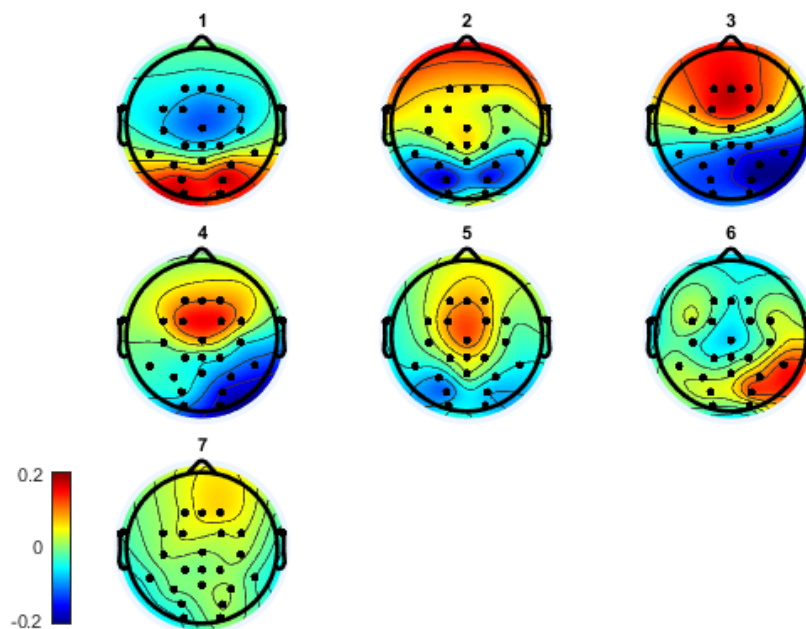


Fig. 3.39 Grand average patterns from each 50 ms long time window from 150 to 500 ms of the group classifier trained on data cleaned from eye activity. The calibration data was all data from recording session 1 and 2 and the classifier was trained on the 24 electrode setup.

3.2.6.2 Discussion

In this part of the study, we decreased the number of the initial 50 EEG electrode setup. We first symmetrically removed electrodes from positions between other electrodes, recording mostly redundant EEG activity. We then removed the remaining frontal electrodes as the error-detection classifier we developed so far in parts 1 to 5 of this study, is not calibrated on eye related activity as we removed eye activity from the calibration data through ICA.

Using all data recorded in the first and second data recording session of this study, we first calibrated subject-specific classifier models. Statistical testing revealed, that the number of electrodes in the setup has an effect on the yielded calibration error rates of these subject-specific classifiers. Both, the setup with 32 and the one with 24 electrodes had significantly higher calibration error rates, than subject-specific classifiers trained on all 50 electrodes. Though this difference is statistically significant, it should be noted, that the average difference between calibration errors of the 50 electrode setup ($M=7.27$ percent) and both, the 32 electrode setup ($M=8.34$ percent) and 24 electrode setup ($M=8.37$ percent) is only a little more than 1 percent, which in practice will most likely not have a noticeable effect on classifier application accuracy. Calibration error rates are not statistically different between the 32 and the 24 electrode setup. The normalized grand average patterns of subject-specific classifier models have almost identical time courses and topographies regarding the classifier weights in all three electrode setups.

When training the group classifier on all data recorded in the first and second data recording session, classifier error rates improve minimally from the 50 (17.36 percent) electrode setup to the 32 (17.15 percent) and the 24 (17.28 percent) electrode setup, with the lowest calibration error rate in the 32 electrode setup. Again, patterns of the group classifier models have almost identical time courses and topographies regarding the classifier weights in all three electrode setups.

As the calibration accuracy for the subject-specific error-detection classifiers is decreasing with reduced electrode setups, it could be concluded that a reduction of electrodes is inadvisable. However, this decrease is very small (little more than 1 percent on average) which will most likely not have serious effects for classifier application. The slight decrease in calibration accuracy must also be weighed against the advantage which a reduction of electrode numbers in the setup can hold for real-world applicability of BCIs: Application of hardware, the electrode cap in particular, is still one of the big bottlenecks for BCI technology preventing its feasible usage in everyday contexts. A potential user will be reluctant to use a system which needs a long preparation time each time they want to use it for support in their everyday tasks. In general, to assure good data quality, each electrode of the EEG

headset needs to be examined (and corrected) for good contact to the user's scalp before recording EEG data. In the case of gel based EEG headsets, a conductive gel must be applied to each electrode. Handling each electrode adds up to long preparation times, unsuitable for real-world contexts. Hence, a reduction of electrodes in the setup is inevitable for BCI systems that should be used in the real world. In this study, we aim to define such an error-detection classifier which can be applied by users in real-world scenarios one day. Thus, and as the difference between the 32 and the 24 electrode setup is not significant for the subject-specific classifiers and for the group classifier, classification accuracies improve from the 50 electrode setup to both other reduced electrode setups, we decided to continue our analysis in the next part of this study with the 24 electrode setup we introduced here.

In the next part of the study, we investigated whether the defined group classifier trained on data from 24 EEG channels is both, session- and subject-independent. We will explore the group classifier's application performance in comparison to subject-specific classifiers trained on subjects' data from a new third recording session. To examine the group classifier's performance in a real-world scenario, subjects in a new recording session, will experience online adaptation and correction through the error-detection BCI model while playing the RLR-game.

3.2.7 Part 7: Session- and Subject independence - Application of the group classifier to data from new subjects from a third recording session

In this last part of the study, we investigated whether the group classifier we defined - based on all data from the first two recording sessions and calibrated on 24 electrodes - is both, subject- and session independent. To this end, we recorded data from new subjects in a third recording session. Then we compared subject-specific classifier performance based on the subject-specific new calibration data to the application performance of the already defined group classifier. Finally, subjects experienced online adaptation of the RLR-game through an error-detection group classifier BCI and we evaluated the group classifier's performance in this real-world online application scenario.

3.2.7.1 Methods

Participants In this third experimental data recording session, data from 16 subjects was recorded (age: 19 to 40 years).

Apparatus EEG data was recorded from 32 active Ag/Cl-electrodes with electrode sites according to the extended 10-20 system and with a biosignal amplifier (Brain-products GmbH, Gilching, Germany). The sampling rate of the recordings was 100 Hz. EMG and EOG were recorded additionally from 4 electrodes. The reference was the nasion (and the ground electrode at Fpz). All electrode impedances were kept below 20k Ω .

Experimental paradigm The experimental paradigm remained the same as in recording sessions 1 and 2. Subjects played the RLR-Game, against an opponent. We recorded data in four blocks: In block 1, subjects played the game, without any errors occurring. In the second block, errors were introduced in 30 percent of trials. In a third block, subjects played the game again, with errors induced, but this time error correction through the error-detection BCI was simulated in real-time, while subjects played the game. In a fourth block, errors were detected and automatically corrected online through the use of a group classifier BCI.

3.2.7.2 Data Analysis

Recording blocks 1 and 2 from the third data recording session served as calibration data for subject-specific classifiers from subjects of this new recording session. We first cleaned the data from eye activity. Again, we computed an ICA for each subject and used the IClab (Pion-Tonachini et al., 2019) plugin to automatically classify

EEG independent components. Independent components identified as eye activity with at least 50 percent certainty, were subtracted from the respective subject's data.

We then plotted grand average ERPs of these datasets. First, we filtered the data by applying a lowpass filter of 40 Hz and a highpass filter of 0.5 Hz. Data was rereferenced to the common average using the fullRankAveRef (Miyakoshi, 2017) EEGLAB plugin. Epochs of 1.2 seconds length were extracted starting 200 ms before erroneous and correct rotations. A baseline correction was performed using the 200 ms before stimulus onset. Grand average ERPs were generated from all trials of all participants for erroneous and correct rotations. The resulting grand averages and difference curves were plotted for electrode sites Fz, Cz, and Pz as in part 1 and 2 of this study (see section 3.2.1.2).

Then we calibrated subject-specific classifier models with the same 24 electrode setup as the group classifier we defined in part 6 of this study on datasets 1 and 2 from each subject of this third session. Here we followed the same classifier calibration approach as described in part 1 of this study (see section 3.2.1.2). We plotted normalized grand average patterns of the subject-specific classifiers. Further, we again analysed weighted dipole density plots showing the relevance of cortical areas to the subject-specific windowed means classifiers. We then applied the obtained subject-specific classifier models, as well as the group classifier trained on 24 electrodes, to data of recording block 3 of the respective subject, yielding application error rates for both classifier types per subject. Finally, we analysed data from recording block 4, in which the game paradigm was adapted online through an online group classifier while subjects played the game.

3.2.7.3 Results

In this section, we present the results of the ERP analysis, classifier calibration and classifier visualisation from the third data recording session. Results from subject-specific classifier calibration will be presented, followed by a comparison of application performance between these subject-specific classifiers and the afore defined group classifier. Finally, results from online performance of the group classifier is reported.

Event-related potentials The grand average ERPs on electrode sites Fz, Cz, and Pz from calibration data of the third recording session, cleaned from eye activity, are depicted in figure 3.40. The respective ERP difference curves (error minus no error) between the two trial classes are shown in figure 3.41. We again analysed the differences between the two trial classes by performing a permutation test for every sample in the ERP with 5000 permutations. P-values of samples with significant

permutation test outcomes are indicated as gray bars in the background.

The time-course and difference curves of the ERPs at all three electrode sites are very similar to those of the calibration data from the first recordings session, which was also cleaned from eye activity.

The difference curve at Fz is similar to the ERP from study part 2, with a positivity that is a little more narrow and significant from 260 to 360 ms, with a peak at around 320 ms. Again a small, non-significant early negativity is present, peaking around 120 ms. Moreover, at Fz another late and broad negative component is present where permutations tests showed the difference to be significant for many samples from 560 to 700 ms.

At electrode Cz no early negativity present. As in study part 2, again a strongly pronounced positivity for erroneous trials with a peak at around 320 ms can be observed. Similar to parts 1 and 2 of this study, this positivity is again highly significant between 270-440 ms. In the difference curve also another late broad significant positivity can be observed from 490 to 560 ms.

At electrode Pz the difference curve again, as in part 2, shows a delayed early negativity with a peak at around 270 ms. This negativity is significant between 260 and 270 ms. This negativity is again followed by a broad positivity that is significant from 370 to 760 ms.

Subject-specific classifier calibration For all 16 participants of recording session 3, we calibrated subject-specific classifier models on the same 24 electrodes as in the group classifier setup. The obtained classification error rates derived from the data cleaned from eye activity are shown in table 3.14. The average classifier calibration error was 9.68 percent (SD=3.61 percent).

Table 3.14 Classifier calibration error rates in percent with standard deviations of the subject-specific classifiers from recording session 3 (calibration data was cleaned from eye activity)

subject	calibration error	SD
1	5.58	3.03
2	10.77	5.03
3	5.81	3.31
4	8.49	5.78
5	13.90	3.03
6	4.47	3.37
7	10.91	4.07
8	15.51	4.91
9	5.36	2.82
10	8.22	3.22
11	10.61	6.68
12	9.06	1.07
13	14.21	2.61
14	7.55	4.81
15	6.31	1.39
16	18.07	2.62
average	9.68	3.61

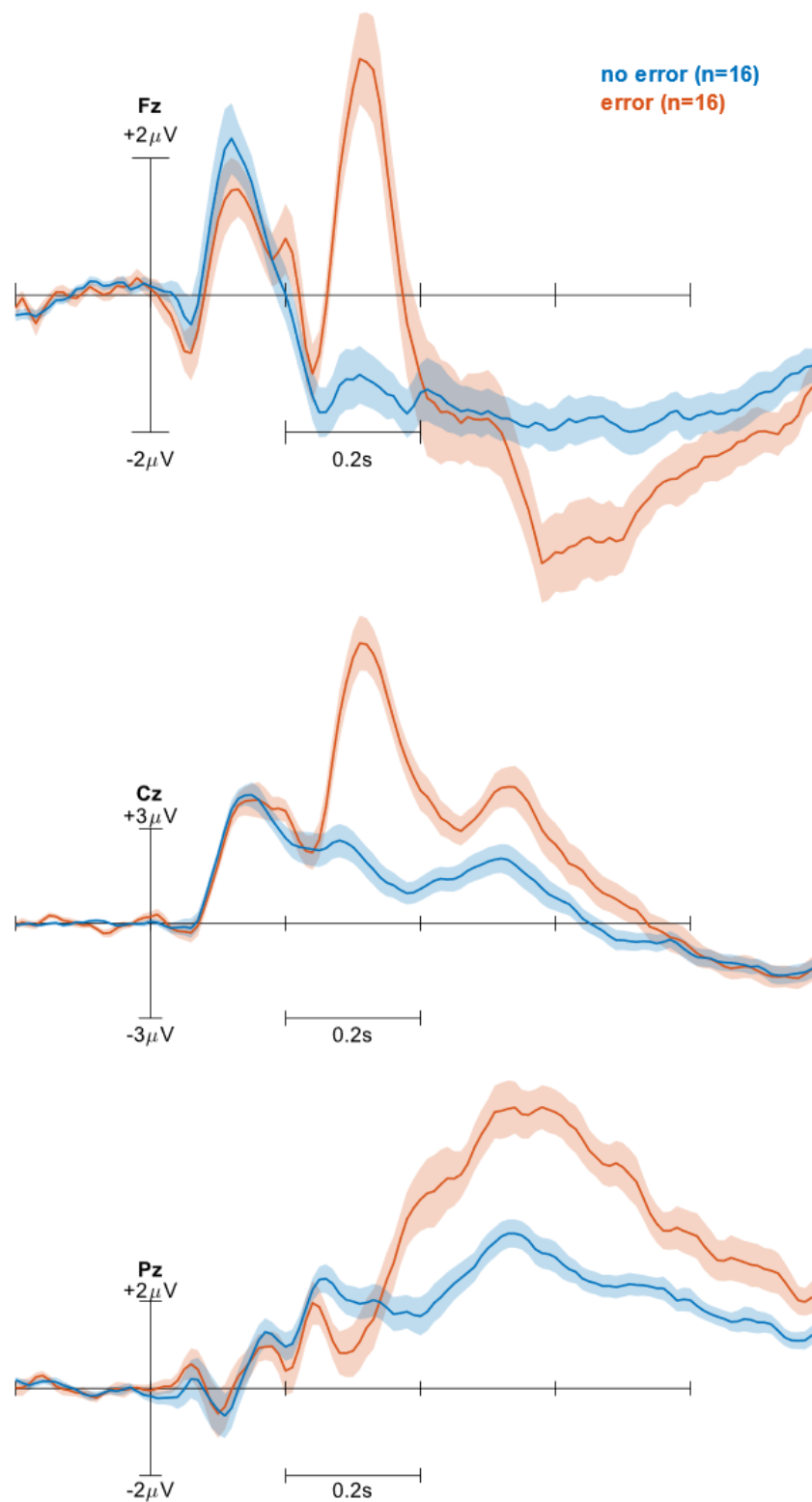


Fig. 3.40 Grand average ERPs (n=16) from data cleaned from eye activity through ICA. Curves show error and no error trials with standard deviations from electrode sites Fz, Cz and Pz from the third recording session of the experiment.

Subject-specific classifier visualisation The normalized grand average patterns from each 50 ms time window from 150 to 500 ms of the subject-specific classifier models of all 16 subjects are shown in figure 3.42. The patterns show that these classifiers have the highest weights in the time window 300-350 ms, where the focus is dipolar on central and occipital sites. This is similar to the subject-specific classifier patterns from recording session 1, though there the occipital sites were more right lateral. The classifier weights in this recording session are also more and mainly focused on the time-window 300-350 ms, while in recording session 1 this was also the time-window with the most weights, but other time-windows received significant classifier weights as well. The weighted dipole density plots of the subject-specific classifiers from the third recording session are shown in figures 3.43 to 3.49.

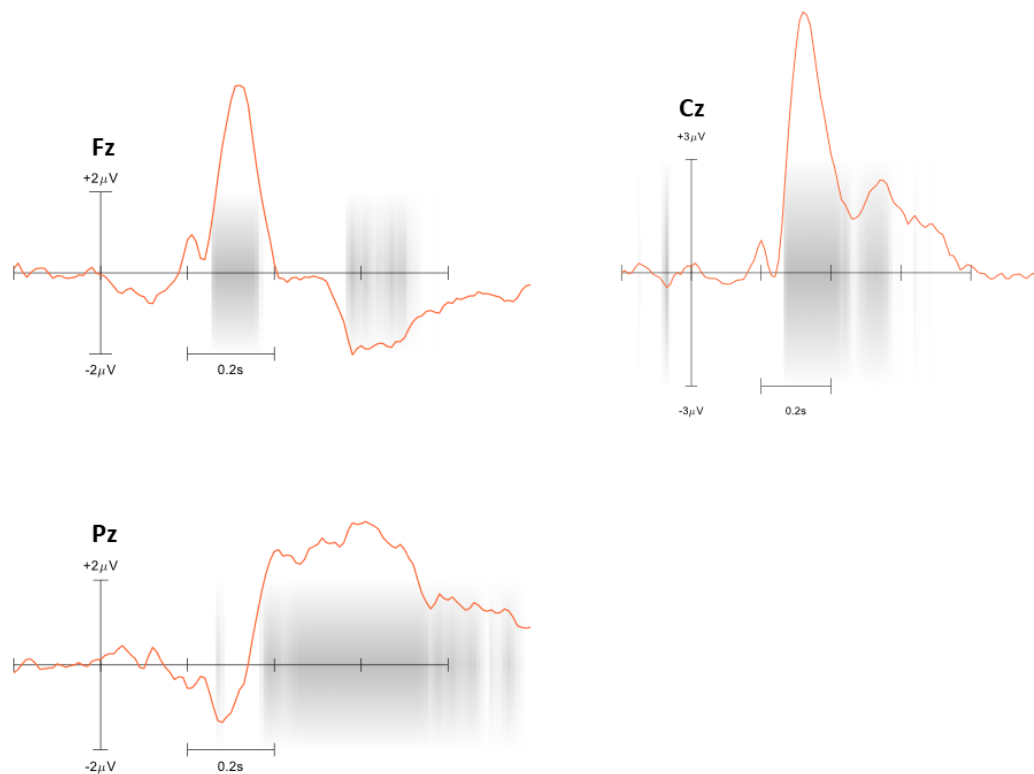


Fig. 3.41 Grand average difference curves of the ERP (error minus no error) of data cleaned from eye activity through ICA of the third recording session at electrode sites Fz, Cz and Pz. Gray bars behind the plot indicate significant p-values from the permutation tests that were performed for each sample.

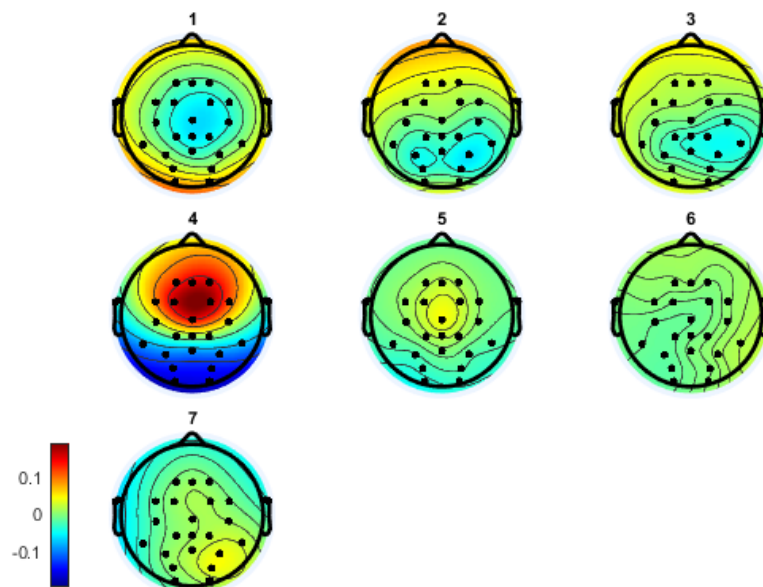


Fig. 3.42 Grand average patterns from each 50 ms long time window from 150 to 500 ms of the subject-specific classifier models of all 16 subjects from the third recording session trained on data cleaned from eye activity.

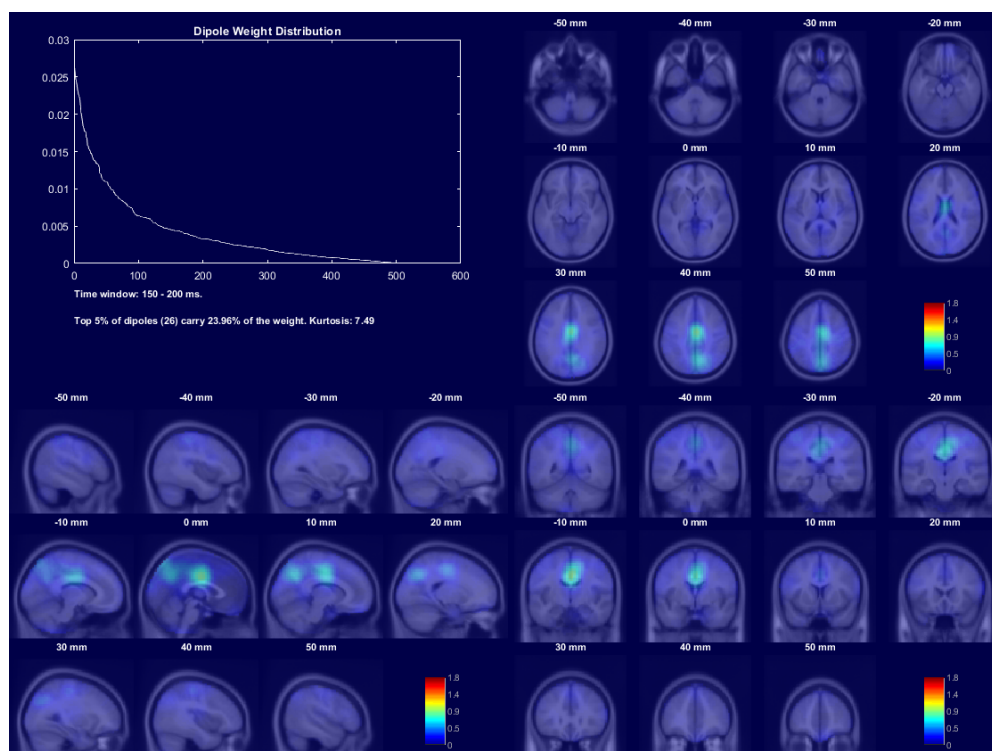


Fig. 3.43 Visualisation of the subject-specific classifiers from recording session 3 trained on data cleaned from eye component activation in time window 150-200 ms: Weighted dipole density plots showing the relevance of cortical areas to the windowed means classifier. Slices are labelled with their corresponding MNI coordinates. Top left: sorted dipole weight distribution.

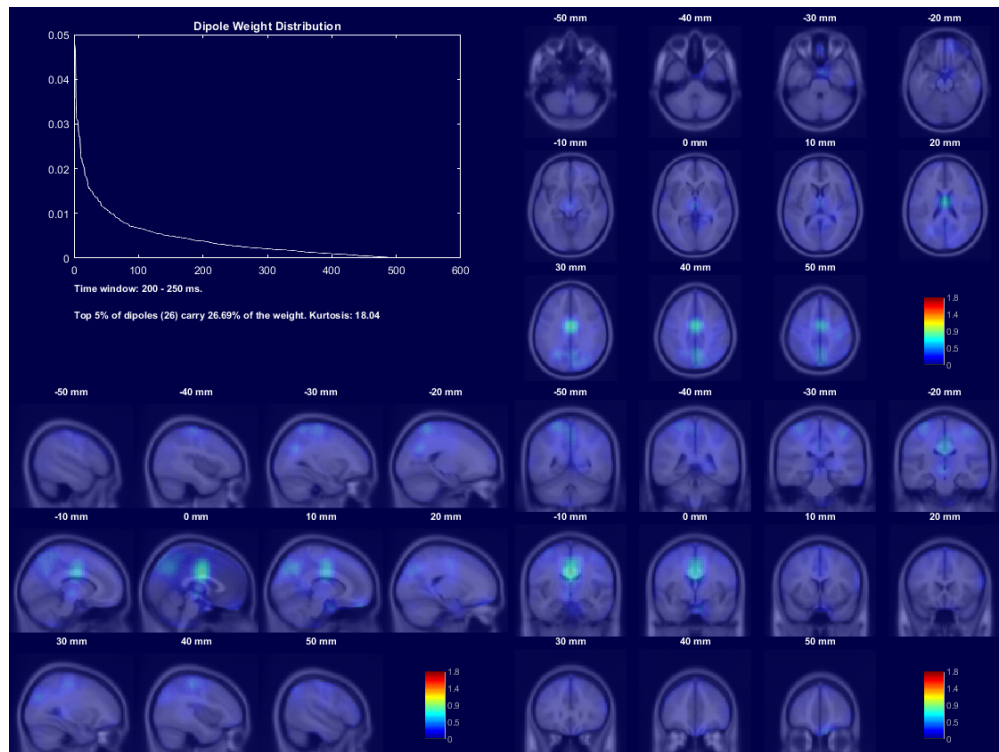


Fig. 3.44 Visualisation of the subject-specific classifiers from recording session 3 trained on data cleaned from eye component activation in time window 200-250 ms.

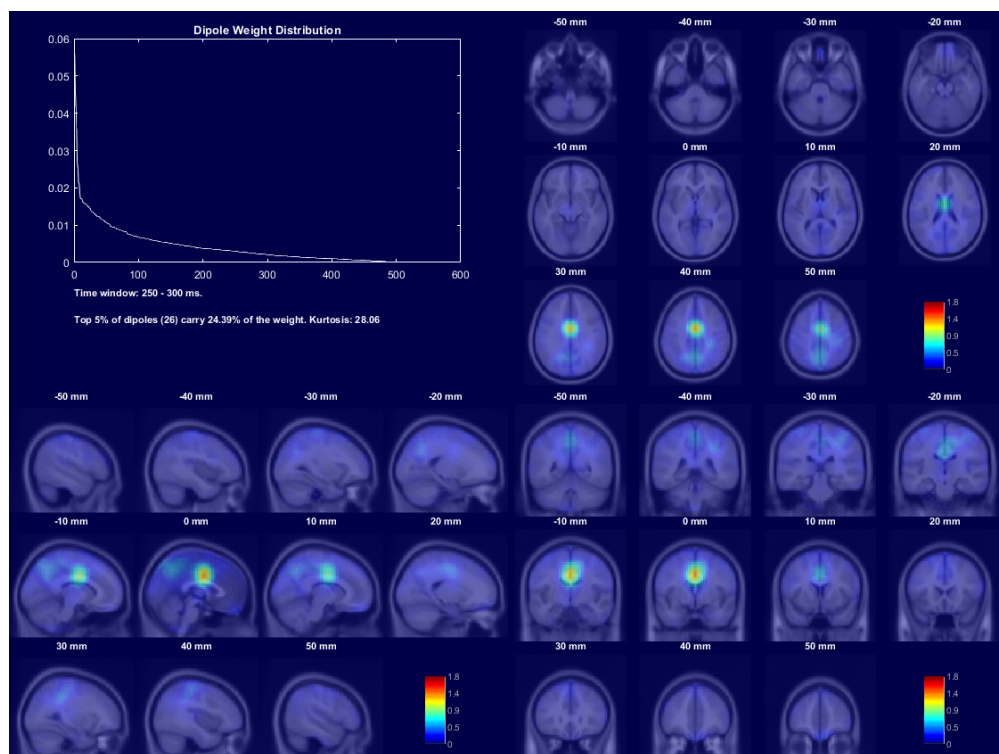


Fig. 3.45 Visualisation of the subject-specific classifiers from recording session 3 in time window 250-300 ms.

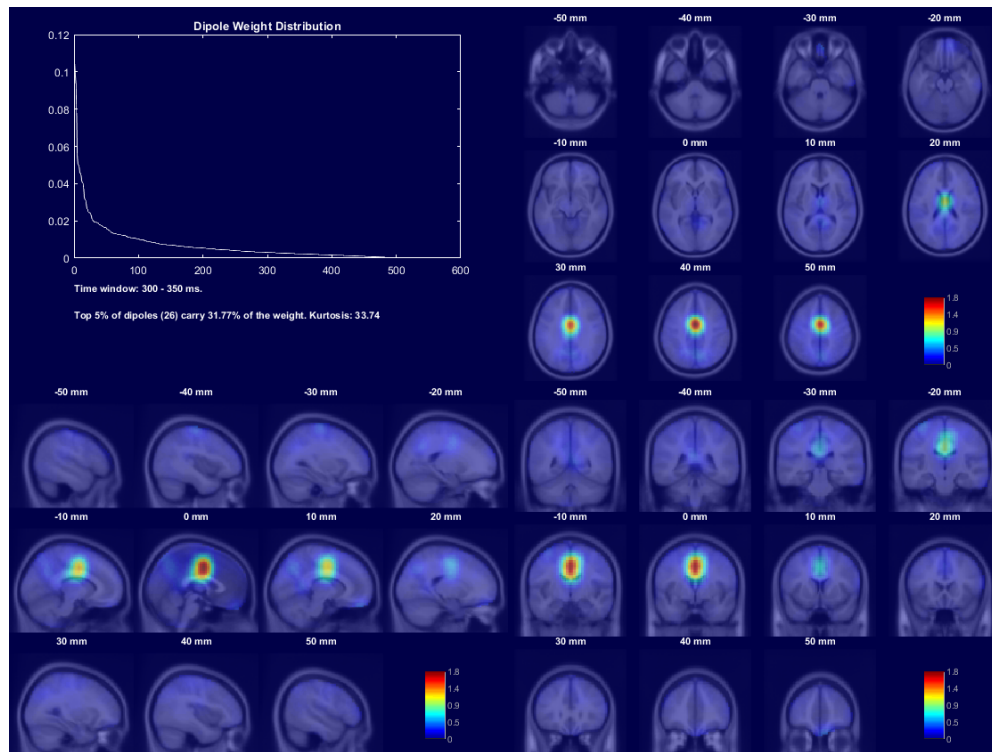


Fig. 3.46 Visualisation of the subject-specific classifiers from recording session 3 in time window 300-350 ms.

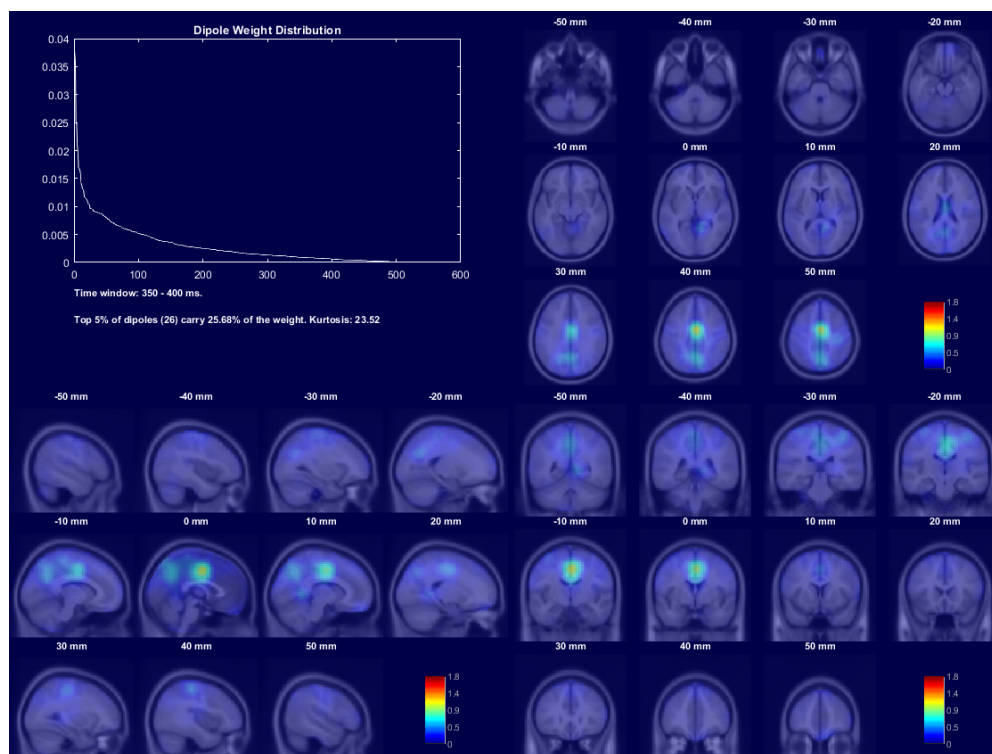


Fig. 3.47 Visualisation of the subject-specific classifiers from recording session 3 in time window 350-400 ms.

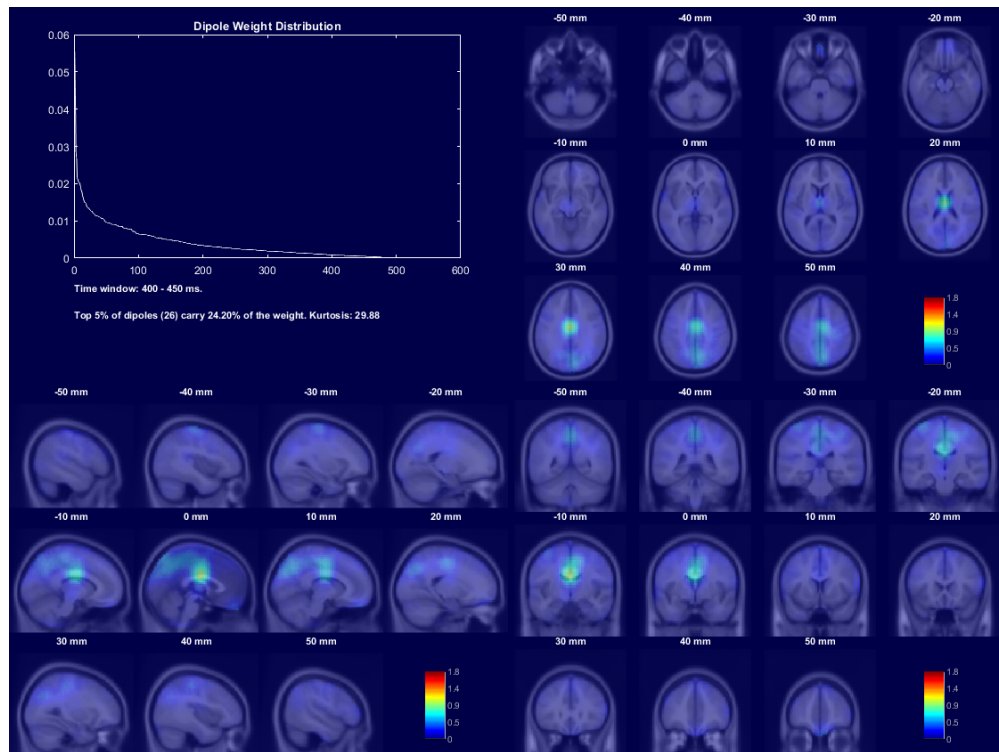


Fig. 3.48 Visualisation of the subject-specific classifiers from recording session 3 in time window 400-450 ms.

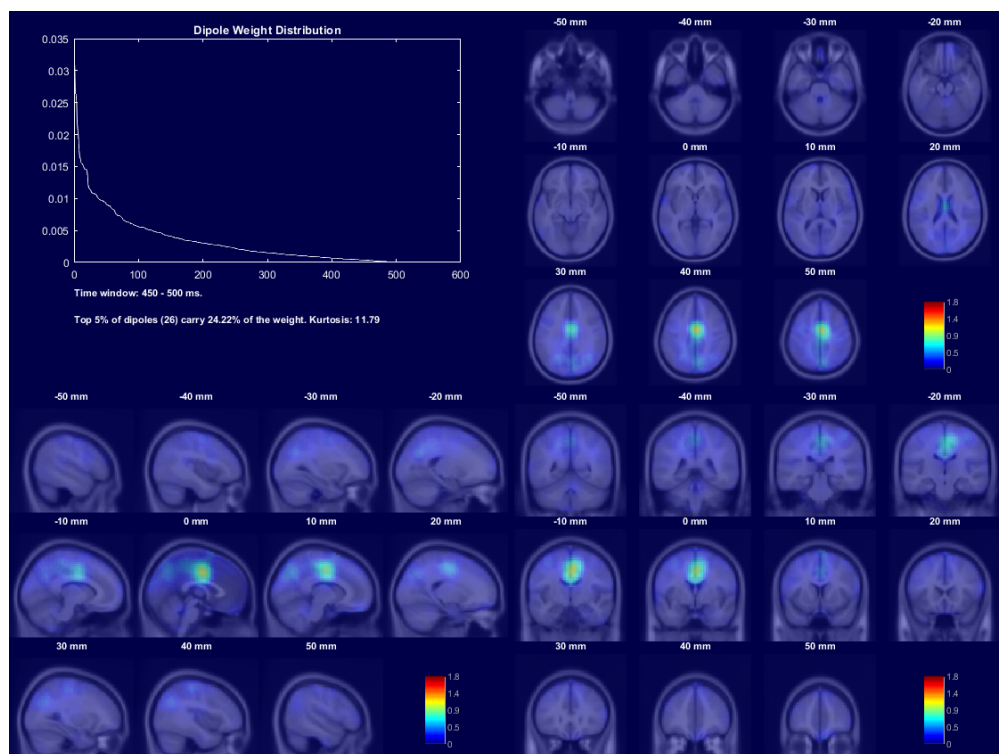


Fig. 3.49 Visualisation of the subject-specific classifiers from recording session 3 in time window 450-500 ms.

Subject-specific classifier of recording session 3 and group classifier application

We applied both, the subject-specific classifier models calibrated on data from recording session 3 and the group classifier (trained on data from recording sessions 1 and 2) to data of the third recording block of recording session 3, in which subjects played the RLR-Game with induced errors and a simulated online error-correction through BCI support. The application error rates for both classifier types are shown in table 3.15. We subjected these application error rates of both classifier types to a paired t-test. Results showed, the difference between application error rates of subject-specific classifiers (M=20.15 percent, SD=7.38 percent) and the group classifier (M=23.78 percent, SD=7.94 percent) is not statistically significant, $t(15)=2.097$, $p=0.053$.

Table 3.15 Classifier prediction error rates in percent of applying the subject-specific classifiers from recording session 3 and the group classifier, both trained on 24 electrodes, to subject-specific testdata of recording session 3.

	subject-specific classifiers	group classifier
subject	prediction error	prediction error
1	14.47	25.79
2	31.00	18.00
3	9.82	16.96
4	15.61	18.05
5	14.81	29.01
6	19.71	16.79
7	26.83	25.85
8	32.21	30.87
9	7.81	9.38
10	14.85	20.30
11	23.47	30.52
12	24.24	26.77
13	21.29	19.80
14	23.13	29.38
15	14.52	20.16
16	28.57	42.86
average	20.15	23.78

Performance of the group classifier in an online scenario Finally, we analysed, how well the group classifier trained on 24 electrodes performed in an online application setting, where subjects received real BCI support through an error-detection

group classifier. The resulting error rates of the group classifier model for each of the 16 subjects is shown in table 3.16. The average error rate of the online application of the group classifier was at 22.76 percent (SD=8.18 percent).

Table 3.16 Group classifier error rates in percent from the online application scenario of recording session 3

subject	error rate
1	16.34
2	20.00
3	13.43
4	15.34
5	24.08
6	12.80
7	29.47
8	37.13
9	16.13
10	20.87
11	20.73
12	30.60
13	20.83
14	26.97
15	18.98
16	40.43
average	22.76

3.2.7.4 Discussion

In this final part of the study we investigated, if the group classifier, which we defined on data from all subjects of the first two recording sessions and calibrated on 24 electrodes, is subject- as well as session-independent. For this investigation, we recorded calibration- and testdata in a third recording session with new subjects. We calibrated subject-specific classifier models on data from this new recording session. These subject-specific classifiers were similar to subject-specific classifiers trained on data from the first recording session in terms of calibration accuracy, pattern features and the visualisation in weighted dipole density plots showing the relevance of cortical areas to the windowed means classifiers.

We then compared classifier application performance between these subject-specific classifiers and the group classifier, when they are applied to new data. In the experimental paradigm for the testdata recordings, an online error-detection

and correction support through the BCI was simulated, in order to achieve a more realistic scenario of the real-world application of such error-detection support BCI for the user. Results showed, that there was no significant difference between the classifier application performance of the subject-specific classifiers and the group classifier. This means, that we can train a group classifier on data from other subjects from different recording sessions and yield similar classifier application accuracies, as when we train a subject-specific classifier model on data which was recorded in the same session. For real-world application of an error-detection classifier this means, that the application of a group classifier is preferable, because it eliminates the collection of calibration data and the calibration phase itself. The support through the error-detection BCI becomes a plug-and-play technology, which can be made use of immediately after the EEG headset is mounted to the user's head.

Finally, all subjects in this third recording session, experienced real online-adaptation and error-correction support through the BCI classifier model of a group classifier. Here we examined the group classifier's performance in such a real-world online scenario. The results showed, that in this online application, error rates of the group classifier were comparable to the simulated online experience discussed before. We conclude that the group classifier which was trained on 24 electrodes of data from other recording sessions (all data cleaned from eye component activity), is indeed subject- and session-independent and can be successfully and reliably applied in an online application scenario.

3.3 General discussion

In the presented study and its seven distinct parts of analyses, we gradually examined whether an error-detection passive BCI classifier model can be trained in a way that it is session-independent, subject-independent and both at the same time for calibration time reduction. We also tested the applicability of such a session- and subject independent error-detection classifier model in an online real-world scenario.

We first trained subject-specific classifier models based on approaches for error-detection classifier calibration deducted from BCI literature. We saw that in erroneous trials our experimental paradigm indeed evokes ERP components commonly associated with error-processing, namely an ERN followed by a Pe. We trained classifier models on subject-specific data, following a windowed-means approach, and applied these classifier models to subject-specific testing data. Resulting error rates were low for classifier calibration and application. Classifier visualisation however showed that the classifier models were basing their decisions not on activity from cortical areas, but focused on eye activity. Hence, we first tested if one could calibrate and apply classifiers on EOG data only. We found that classification in this case was not fully random, meaning that eye activity holds some information about error processing, but application accuracies were much too low for meaningful classifier application in real-world scenarios.

In a next step, we excluded eye activity from the calibration data to calibrate error-detection classifier models which would not base their decisions on eye activity, but on differences in cortical activity. These subject-specific classifiers trained on eye activity free calibration data had low error rates for classifier calibration and slightly increased classifier application error rates, which were on average still below 20 percent. Classifier visualisation showed, that these new, now truly "brain"-computer interface models, were basing their decisions mainly on a cortical area around the anterior cingulate cortex (ACC) which neuroscientifically is associated with error-processing. The time-window holding the most relevance for classification was from 300-350 ms after stimulus onset. The subject-specific models from this part of the study still required the recording of new calibration data and calibration time to train a new classifier model in each session, when the specific subject wants to use the error-detection classifier.

Therefore, in the following part of the study, we explored whether session to session calibration of subject-specific classifier models could be omitted - if the error-detection classifier was *session-independent*. To this end we recorded new testdata from the same subjects on a different day and applied the already calibrated subject-specific error-detection classifier models to these new testdata. Results

showed, that indeed classifier application error rates (with an average of around 20 percent) were not different from error rates from classifier application within the same recording session. These results demonstrate, that intra-individual error-processing is a process in the individual human's brain which is constant over time and we can indeed train an error-detection classifier, which is session-independent. For real-world application this means, that the calibration data for a subject-specific classifier only must be recorded once for an initial classifier calibration and that the user could from then on use this classifier model immediately in a new session for online support through the BCI.

As we had calibration data from more than one subject available, we investigated the question, whether a classifier trained on the specific subject's data, as well as calibration data from all other subjects would be session-independent as well - by yielding classifier calibration and application error rates similar to the subject-specific classifier models. We also wanted to investigate, if an error-detection classifier trained on calibration data from multiple subjects would base its decisions on activity originating from the same cortical areas as the subject-specific classifier models. We trained this *group classifier* on calibration data from all subjects of the first recording session and applied it to the subject-specific testdata from the second recording session. Results showed, that the calibration accuracy of the group classifier was slightly lower than in subject-specific classifier models, but that the classifier application accuracy of the group classifier was not statistically different from the performance of the subject-specific classifiers. Classifier visualisation showed, that the group classifier bases its decisions on very similar cortical areas and time-windows as the subject-specific classifier models. These results show, that on data from more than one subject, we can train a functioning error-detection classifier model as well, which is working on the same neural basis as subject-specifically trained classifiers.

Through both, the session-independent subject-specific and the group classifier, calibration time for online-application of the error-detection classifier model can be substantially reduced - from recording new calibration data in every application session to just one time in an initial session. In order to supersede calibration time for an individual user completely, one could also try to calibrate the error-detection classifier model on completely different people's data and apply it to testdata of the specific user. Hence, we tested whether our error-detection classifier could work in this sense of *subject-independence*. To that end, for each subject we used the subject-specific classifier model of the subject and applied it to testdata from all other subjects within the first data recording session. Results showed that, at a classifier application accuracy of around 67 percent, the subject-specific classifier application to data from other subjects was much worse than when the

respective subject's own classifier model was applied (here application accuracy was about 81 percent). We also calibrated group classifiers in terms of leave-one-subject-out calibration. Here, for each subject, we calibrated the classifier model on data from all other subjects and then applied this group classifier to testdata of the specific subject, whose data was not part of the calibration data. The resulting leave-one-subject-out group classifier application accuracies were on average at 73 percent, which again is significantly worse than when the respective subject's own classifier model was applied (with an application accuracy of 81 percent). However, the group classifier's performance also was significantly better, than when the subject-specific classifier models were applied to other subjects' data. This investigation into subject-independence of the error-detection classifier showed, that we can apply subject-specific, as well as group classifiers to data from a completely new person. Though both classifier types yield worse classifier application performance, than a classifier specifically trained on the specific subject's calibration data, their application error rates are far from random, showing that the neurophysiology of error-perception is inter-individually highly consistent. The subject-independent group classifier's application performance is significantly better than when classifier models calibrated on data from one specific person are applied to data from a different person. Even though there is an average decrease of 7 percent in application performance from the group classifier to the classifier model trained on the specific person, the subject-independence of the group classifier would not require any calibration time in any recording session from the end-user of the BCI system. Classifier calibration would be fully omitted, because the error-detection classifier model can be trained beforehand on data samples from other individual users. As calibration time is one of the critical bottlenecks for BCI technology to find its application in everyday HMI contexts, the benefits of subject-independence may outweigh the slight decrease in classifier application accuracy.

After finding in the previous parts of the study, that we can train an error-detection classifier model, which works session- or subject independent, in the final part of the study, we aimed at combining both features - session- and subject-independence - in the error-detection classifier model. As classifier application performance in the subject-independent classifier model was significantly better for the group classifier, we decided to calibrate a new group classifier on all available data from all subjects from the first two recording sessions, we then applied this group classifier model to data from completely new participants of a new recording session. One step before training this group classifier, we evaluated, whether the number of electrodes (we train the error-detection classifier model on) could be reduced. The reduction of required electrodes is another crucial factor, when it comes to real-world applicability of BCI technology. Every electrode that must be applied to the scalp

and adjusted for good skin contact, adds to the time required to set up the EEG system, before it can actually be applied in a given task. We first reduced the initial 50 electrode setup down to 32 and then finally to 24 electrodes. Classifier calibration performance of subject-specific classifiers showed that there was a statistically significant but practically dismissable (on average 1 percent) decrease in calibration accuracy from the 50 electrode setup to both, the 32 and the 24 electrode setup. There was no difference found between the 32 and the 24 channel setup. Calibration accuracy of the group classifier model increased slightly from 50 electrodes to both, the 32 and 24 electrode setups. These results indicated that we could reduce the number of electrodes we would calibrate the group classifier on down to 24.

In the final part of the study, we recorded new calibration and testdata from new subjects. We used the calibration data to calibrate new subject-specific classifier models on 24 channels for each of the subjects. After analysis and visualisation of these new subject-specific classifier models, we applied them to the respective subject's specific test data. The resulting application error rates were compared to error rates of the group classifier to the same testdata. Results showed, that there was no difference found between classifier application performance of the specifically trained classifier models and the newly defined group classifier. Finally, we tested group classifier performance in an online feedback scenario, where automatic error correction was applied while subjects played the experimental paradigm game. Here the online application accuracy of the group classifier was on average at 77 percent. The results from the final part of our study showed, that for automatic error-detection we can train a group classifier, on calibration data of multiple subjects, with a reduced electrode number of 24 channels. This group classifier is session- and subject-independent, as it can be applied to new data with similar application accuracies as a specifically trained error-classification model. These application accuracies remain stable even in an online application context, where the classifier is used for real-time detection of error potentials and automatic error correction while the user performs their task.

3.4 Conclusion

There are several factors today which prevent neuroadaptive technology which is enabled by passive BCIs from real-world application by users in their everyday lives and professional environments. One of these factors is the calibration time needed to collect training data and calibrate a classifier model for the brain-computer interface's classifier model. Another factor is the time, effort and expertise needed to mount an EEG headset to the user's head before data recording and monitoring can begin - this factor depends strongly on the number and type of EEG electrodes that are needed for the application.

In this study, we investigated ways to reduce both - classifier calibration time and EEG headset preparation time. For an automatic error-detection classifier - one of the most promising and meaningful passive BCI applications for HMI contexts in the future - we showed, that a classifier model can be trained which omits all classifier calibration time for the user. We also demonstrated that the number of electrodes this classifier model is trained on, can at least be halved from the initial 50 electrodes used in our study down to 24 electrodes. These two improvements to system preparation time bring automatic error-correction support through passive BCI support one step closer to be applied in the real-world.

Above that, we also assured thorough classifier model visualisation and validation, that the classifier models we calibrated and applied were truly basing their decisions on cortical sources in the brain and not on changes in electrical activity coming from other sources or modalities, such as the eyes or muscle movements.

The results of our study further gave valuable insight into the neurophysiology of error processing. They suggest, that error processing in the human brain, even in single-trial, is a highly similar process in humans. We demonstrated that inter- and intraindividual variance in neurophysiological changes during error processing does not vary much, making error processing a strongly constant brain process across subjects.

Regarding the grading of the demonstrated classifier application accuracies, the actual benefit of using an error-detection passive BCI for automatic error correction in an HMI context is highly dependent on the situational context, in terms of time, frequency of the application and the impact of errors made in the specific context. For example, if an automatic error-detection BCI is used to revert and correct auto-corrections made by the program during email writing, it is not highly important, that every error in the auto-correction is detected and adapted. Hence, in this context the demonstrated classifier application accuracies will most likely be sufficient. But there might also be other contexts, where higher accuracies could be crucial, e.g. in high focus and danger situations. However, automatic error detection will

likely improve system performance even with lower classifier application accuracies, as it even then would detect and adapt some of the errors committed by the system in the HMI context which without the error-detection classifier would not be corrected at all.

Further, more precise performance metrics, such as *balanced accuracy* as a performance measure in unbalanced class classification, should be employed when evaluating the real-world applicability of the defined error-detection classifier. Using balanced accuracy or another measure that takes into account class imbalance would allow for more precise statements regarding whether the classifier truly performs better than the random level.

For future research, the developed session- and subject-independent group classifier should be tested and applied for error detection in further tasks and scenarios in HMI where execution or interaction errors can occur. The demonstrated well-performing classifier model could perform differently under different tasks and contexts. If the here developed error detection passive BCI model could be successfully applied to different tasks for error identification and correction, the classifier would not only be session- and subject-independent, but also task-independent - making it a generally *universal* BCI classifier model that can be applied in a variety of interactions between a human and a system to automatically correct interaction or execution errors without any required calibration time.

Above that, it should also be investigated, whether the developed group classifier model can also detect other types of errors than interaction or execution errors, as e.g. an observation error. Finally, the group classifier should be tested with other EEG headsets (e.g. dry electrodes), as well as with even more reduced numbers of electrodes for its application - as in the context of a real-world application, not all users will have the same EEG headset available as we used to record the EEG data in this experiment.

The willingness of users to apply a BCI for support will however always depend at least on the factors, which form the following formula:

$$\text{BCI application acceptance} = \frac{\text{benefit within context}}{\text{EEG system setup time} + \text{calibration time}}$$

As long as the benefit of the BCI application is greater than the perceived effort the user must put into (repeated) classifier calibration and mounting the EEG headset, users will tend towards application of the BCI support in the real-world. Today however, in most cases, the result is the other way around and it is too cumbersome to prepare the BCI system before starting the task it should support the user in. I want to note however, that the factors determining the acceptability of BCI systems in general and for specific applications and user groups are complex and still un-

der investigation. In a detailed and recent study e.g. Grevet et al. (2023) investigated factors influencing the acceptability of active motor imagery-based BCI systems for post-stroke rehabilitation and showed that system characteristics, especially scientific relevance and benefits on risk ratio seem to determine the *perceived usefulness* of the BCI. Further, facilitating conditions, specifically ease of learning and playfulness, influence the *perceived ease of use* of a BCI system. Finally, they showed that social influence also plays a role.

With the presented study and its results, we have taken a big step towards minimising system preparation time for online application of an automatic error-detection BCI. We hope that this step will bring this application of neuroadaptive technology one step closer to be incorporated in HMI contexts and it will be integrated in real-world applications to support users in fulfilment of their tasks and wishes.

Study 2: Task-independent mental workload assessment in a speed-reading context

Preliminary note: The following study was already published in Andreessen et al. (2021).

4.1 Summary of the study

As introduced in section 2.7.6 the development of task-independent predictive models for specific user states would be strongly beneficial for real-world application of passive BCI technology. As some tasks potentially are based on similar cognitive processes, it seems unnecessarily effortful to design a specific calibration paradigm for each use case or specific task that is evoking the same cognitive process or state across tasks. Krol et al. (2016) proposed a classifier model that is trained to classify examples of mental workload on the basis of data collected from a subject performing an arithmetic task, while this model is intended to quantify mental workload in a generalized way, across tasks. It thus can reliably distinguish between phases of low and high workload, not only in arithmetic tasks but also in other tasks and domains and is therefore referred to as a task-independent predictive model.

In this study, to further test the task-independence of this promising task-independent predictive model for mental workload detection, we applied it to a new task - a speed reading context, where subjects read texts of different difficulty and at different reading speeds. Firstly, subject-specific classifier models were trained after subjects completed this potentially task-independent mental workload calibration paradigm. The resulting predictive models were then applied to monitor mental workload while subjects read text material of different difficulty and reading-speed. We recorded EEG data from each subject, while twelve texts in blocks

of three were presented to them word by word. Half of the texts were easy, and the other half were difficult texts according to classic reading formulas. From each text category three texts were read at a self-adjusted comfortable presentation speed and the other three at an increased speed. For each subject we applied the trained predictive model for workload detection to EEG data of each word of the twelve texts.

We found that the resulting predictive values for mental workload were higher for difficult texts than for easy texts. Predictive values from texts presented at an increased speed were also higher than for those presented at a normal self-adjusted speed. These results suggest that the mental workload based predictive model can be applied to the new tasks of reading texts at different difficulties and reading speeds, supporting the claim that it can indeed serve as a task-independent classifier for mental workload. Such a model could be employed on a single-subject level in a system which continuously monitors brain activity related to mental workload and adapts to specific reader's abilities and characteristics by adjusting the difficulty of text materials and the way it is presented to the reader in real time. A neuroadaptive system like this could foster efficient reading and text-based learning by keeping readers' mental workload levels at an individually optimal level.

4.2 Introduction and motivation

The emerging field of neuroadaptive technologies aims to automatically adapt a machine to changes in its operator's cognitive or affective state through an ongoing evaluation of neurophysiological activity (Krol and Zander, 2017). One tool to assess information about the current mental state of a person is a passive Brain-Computer Interface (pBCI, Zander and Kothe (2011)). A pBCI is trained to distinguish between different cognitive or affective aspects of user state. Therefore, it learns to classify certain patterns of brain activity (i.e. *classes*), typically recorded through an electroencephalogram (EEG) in a controlled calibration session. An approach combining machine learning and signal filtering is used to automatically analyze a segment of EEG and to provide a guess (i.e. through interpreting the classifier output as *predictive values*) to which class the recorded activity belongs (Zander et al., 2010), resulting in a so-called *predictive model*. This model can then be applied to new data providing predictive values and hence information about the current state of the investigated cognitive process. By combining this information over time with contextual information, a predictive user model can be derived that provides meaningful details about certain aspects of higher cognition (Zander et al., 2016). This user model can then be utilized to adapt a system's behavior in a certain situation to the current state or specific aims of the user – the system becomes *neu-*

roadaptive. This neuroadaptation is intended to improve the interaction between user and machine by making it safer, more pleasant or more efficient (Hettinger et al., 2003; Zander, 2012; Zander et al., 2016). A pBCI differs significantly from classic BCIs (i.e. active or reactive BCIs) which are built on brain activity that is generated by the operator intentionally in order to send specific control signals to a computer system, thereby replacing other means of input (e.g. keyboard) (Zander et al., 2010).

4.2.1 Task-independent classification of spectral changes through mental workload

One important cognitive state that might be used for neuroadaptive technologies is *mental workload*, which is a concept that is well established in experimental Psychology, and Human Factors/Ergonomics (Kramer, 2020; Wickens, 2008); and, that is also characterized by specific neurophysiological signatures (e.g., Gevins and Smith (2000); Gevins et al. (1997); Krause et al. (2000); Pesonen et al. (2007); Scharinger et al. (2015, 2017)). Recent studies claim that pBCIs can be trained to distinguish between low and high mental workload states of a user and predict mental workload in different task domains such as e.g. n-back tasks (Grimes et al., 2008; Herff et al., 2014), span tasks (Chaouachi et al., 2011), Sternberg tasks (Baldwin and Penaranda, 2012; Roy et al., 2016), arithmetic tasks (Walter et al., 2017) and visual-spatial task (e.g. mental rotation; Berka et al. (2007)). However, these attempts have often been of restricted success due to several methodological limitations (cf. Gerjets et al. (2014)). Specifically, efforts taken for designing a *task-independent predictive model* that can identify levels of mental workload in general, independent of the task evoking the mental activity, were not sufficiently successful. However, this is important for detecting mental workload in real-world applications, as tasks at hand can change over time. Recently, a more promising task-independent predictive model for mental workload detection has been proposed by Krol et al. (2016). This model is trained on data collected from a subject performing an arithmetic task triggering brain activity that is related to mental workload in general. While it is trained to classify these examples of mental workload, this model is intended to quantify mental workload in a generalized way, across tasks. This goal was not achieved by adapting feature extraction or classification algorithms in the predictive model, but by designing an experimental paradigm for the calibration session. The resulting predictive model was successfully applied to tasks in other domains, such as n-back tasks, backward span task, solving anagrams, mental rotation or understanding verbal descriptions of a complex context (Krol et al., 2016; Zhang et al., 2018). It, thus, can reliably distinguish between phases of low and high workload, not only in arithmetic tasks but also in other tasks and is therefore referred to

as a task-independent predictive model. In this paper, we investigate whether this task-independent predictive model might also be suitable for a passive BCI-based assessment of mental workload imposed by text difficulty and presentation speed during reading. Workload assessment during reading in real time is a crucial prerequisite for the development of neuroadaptive support technologies that aim at improving reading efficiency in the digital age.

4.2.2 Mental workload and text comprehension

Reading is part of our everyday lives. Written language plays a crucial role when humans intend to share information and acquire new knowledge. From a cognitive science perspective, mental workload (in terms of working-memory load) is central for reading as reading comprehension is based on a set of processes at different levels, many of which require working-memory resources (Daneman and Carpenter, 1980; Kintsch and Walter Kintsch, 1998; Perfetti, 1985). For instance, there are reading processes at the word level (e.g., activation of word meaning), at the sentence level (e.g., integrating word meanings with the syntactical structure to a coherent sentence), and at the text level (e.g., integrating the read content plus information inferred from prior knowledge into a local and global coherent situation model). Inferences at the text level can be broadly defined as information that is not explicitly stated in the text but generated by the readers themselves from memory or the given information (Tibus et al., 2013). Inferences are crucial for understanding (“the cornerstone of comprehension”, Kendeou et al. (2016)). As they are essential for comprehension, these are generated “online” during the reception process (e.g. Graesser et al. (2007)). All of the above mentioned processes – but particularly inference generation at the text level – might not only impose time demands for processing onto readers but also high levels of mental workload (depending on readers’ level of reading fluency and prior knowledge). Accordingly, the more difficult a text is and the less working-memory capacity and reading time a reader has at his disposal the more likely it is that the comprehension process might be disrupted. According to Cognitive Load Theory, learning and comprehension will break down once the information that needs to be integrated in working memory at one point in time exceeds the available working-memory capacity (Sweller et al., 1998). In line with this assumption, reading research has shown that individual differences in working memory capacity can account for qualitative and quantitative differences among readers in several aspects of language comprehension (Just and Carpenter, 1992). In sum, the comprehension of textual information is influenced by several factors, including features of text materials itself, such as *text difficulty* (e.g., length and complexity of sentences or use of technical terms; cf. DuBay (2004); Friedman

and Hoffman-Goetz (2006)). Additionally, text comprehension strongly depends on characteristics of individual readers, such as their domain-specific prior knowledge (Moravcsik and Kintsch, 1993), their individual working-memory capacity (Just and Carpenter, 1992), their level of reading fluency (Kuhn et al., 2010) or the time invested in reading a text (i.e., reading speed). Furthermore, these factors will interact with each other. For instance, *reading speed* will influence text comprehension performance in relation to the individual reader's reading fluency (Bell, 2001; Wallot et al., 2014). Moreover, text difficulty and domain-specific prior knowledge will interact in affecting mental workload and reading comprehension. For example, the same text will impose more workload onto low prior-knowledge readers than onto high prior-knowledge readers (cf. McNamara et al. (1996); Sweller et al. (1998)).

Therefore, developing adaptive support technologies aiming at improving reading efficiency is not trivial. This is particularly true when trying to base adaptations on measurements of individual readers' domain-specific prior knowledge, reading fluency, and working-memory capacity. An alternative strategy – that we will focus on in this paper – is to base adaptive systems on an online measurement of mental workload. The reasoning for this is that mental workload during reading is directly influenced by prior knowledge, reading fluency, and working-memory capacity. It is also an important factor for successful comprehension processes.

Moreover, mental workload can be used online in order to adapt the presentation of reading materials, e.g. by adjusting text difficulty. Methodologies of computational linguistics allow to automatically assess and label the complexity (i.e. *readability*) of reading materials based on linguistic features. Traditional methods use rather simple text analysis techniques that provide easily computable complexity-indices (e.g., the *Flesch-Reading-Ease* (Flesch, 1948) formula or the *Flesch-Kincaid-Grade-Level* formula (Kincaid et al., 1975) using easy to count text features, like the average sentence and word length). More advanced approaches to readability analysis employ a range of even more sophisticated techniques, including statistical language models (Collins-Thompson and Callan, 2004; Kincaid et al., 1975; Petersen and Ostendorf, 2009; Si and Callan, 2001), machine learning using lexical and syntactic features (Feng, 2010; Schwarm and Ostendorf, 2005), models integrating language-specific morphological features (Dell'Orletta et al., 2011; François and Fairon, 2012; Hancke et al., 2012), and approaches emphasizing a cognitive perspective on coherence/cohesion (esp. Graesser et al. (2014)). These examples indicate that text complexity can be defined (and consequently automatically assessed) with respect to different linguistic readability metrics. These metrics range from a focus on simple linguistic entities (e.g., words) or complex, interrelated linguistic entities (e.g., sentences, paragraphs, or entire documents). An automated adaptation (Zander et al., 2010) of reading materials based on mental workload could there-

fore be carried out by selecting text materials of an appropriate complexity level for the specific reader. This adaptation can not only be used in order to avoid reading materials that are too difficult, but also to avoid ones that are too simple, which can also impair reading comprehension (McNamara et al., 1996). An even more advanced approach would be to use algorithms for automatically quantifying text difficulty in order to correctly adjust the complexity of text materials that impose too much mental workload (Vajjala Balakrishna, 2015; Vajjala and Meurers, 2014).

4.2.3 Rapid serial visual presentation (RSVP) applications for variable text presentation speed

Besides adapting text difficulty to mental workload, an even simpler method for the adaptation of reading materials to readers' current mental workload is to control for reading time investment by varying the speed of text presentation speed. For instance, state-of-the-art speed-reading applications such as the *Spritz*TM speed-reading application (<http://www.spritzinc.com>) use *rapid serial visual presentation* (RSVP) to present text materials word for word with an adjustable presentation speed. Each word of the text is presented successively at a fixed position on the screen (Hedin and Lindgren, 2007). In the Spritz application a certain letter in each word, the optimal recognition point (*ORP*, Benedetto et al. (2015); Hester et al. (2016)), is highlighted in red and also presented at a fixed position to which readers should allocate their visual focus during reading to minimize eye-movements, thereby reading more efficiently. Moreover, while the nominal reading speed in Spritz is set a priori by the reader, the actual reading speed differs from word to word and is influenced by factors such the length of the words and the length of the sentences, as well as by the current state of the reader. As a result, longer words are displayed longer and longer sentences have longer pauses at the end of the sentence (Benedetto et al., 2015). However, up to now, readers have to either manually adjust the nominal presentation speed of a text passage they are reading (according to their current perception of the adequacy of reading speed) or read all text passages at the same speed independent of difference in difficulty. Manual adjustment is cumbersome, requiring readers to pay attention to the current reading speed and perform manual acts to make adjustments. An adaptive system operating on the measurement of mental workload could conceivably change the nominal reading speed of text materials in a seamless way. This would facilitate an efficient reading experience that keeps readers always within an optimal range of mental workload, avoiding possible misapprehensions as well as boredom.

A pivotal prerequisite for the development of such an adaptive system is (of course) an appropriate online mental workload measure. This measure must be

sensitive to variations in mental workload that are resulting from text difficulty or reading speed. Therefore, as a first step towards the development of such a system, we investigate whether a task-independent predictive model for mental workload, which is calibrated individually for any subject, can be used to distinguish between reading materials categorized as difficult or easy according to classic readability metrics. We also investigate if different levels of reading speed induced by slow or fast text presentation can be differentiated through this predictive model.

4.3 Methods

4.3.1 Subjects

Thirteen subjects (one was excluded from analyses – see results section 4.5.1.1) took part in our study, with an average age of 32.6 years (SD = 7.4 years). Eight of them were female. The native language of all subjects was German and their vision was normal or corrected to normal. All subjects were naive to the experiment and none had former experience with speed reading applications. Prior to the experiment, subjects gave their written informed consent to participate in the study and were paid 20 euros for compensation. The experiment was conducted along a guideline ethically approved by the local ethics committee.

4.3.2 Apparatus

We recorded EEG from 64 active electrodes using an EEG system from Brain Products (Brain Amp DC, Brain Products GmbH, Gilching, Germany). Electrodes were positioned according to the extended international 10/20 system, with ground electrode at AFz and reference at FCz, and data was rereferenced to common average reference (CAR). Electrode impedances were kept below 5 k Ω . EEG data was recorded using the Brain Vision RDA (Brain Products GmbH, Gilching, Germany) and the LabRecorder (Delorme et al., 2010). The native sampling rate was 5000 Hz and the recording was then subsampled to 500 Hz. The Matlab embedded toolbox EEGLAB (Delorme and Makeig, 2004) was used for EEG data analysis. All BCI classification and model application was performed using the open source toolbox BCILAB (Kothe and Makeig, 2013).

The experimental paradigm for calibrating the predictive model (*mental workload paradigm*) was implemented using SNAP (Iversen and Makeig, 2019). The *text reading paradigm*, presenting the texts to the subjects, was implemented in Matlab, using the Psychophysics Toolbox extensions (Brainard and Vision, 1997; Kleiner et al., 2007) and the Spritz API. Only the reading panel of the Spritz reading interface was shown and there was no menu to adjust reading speed or to play and pause text presentation. The ORP was highlighted in red as in the original Spritz reading interface (see 4.1 b)).

4.3.3 Stimuli and procedure

Subjects completed both of the above-mentioned experimental paradigms as shown in figure 4.2 (mental workload paradigm and text reading paradigm). The mental workload paradigm (see figure 4.1 a)) served as calibration session for the predictive model, presenting a series of stimuli inducing low and high mental workload.

This paradigm is introduced in Krol et al. (2016) and is described in more detail below. In the text reading paradigm (see figure 4.1 b)), subjects were asked to read twelve texts. Each text was presented at a certain speed (normal and fast) and had a certain difficulty (easy and difficult). For each combination of text difficulty and presentation speed three different texts were presented to the subjects, resulting in an experimental design with 3x4 texts. Details of this experimental paradigm are described below, and the timeline of the full experimental procedure is depicted in Figure 2.

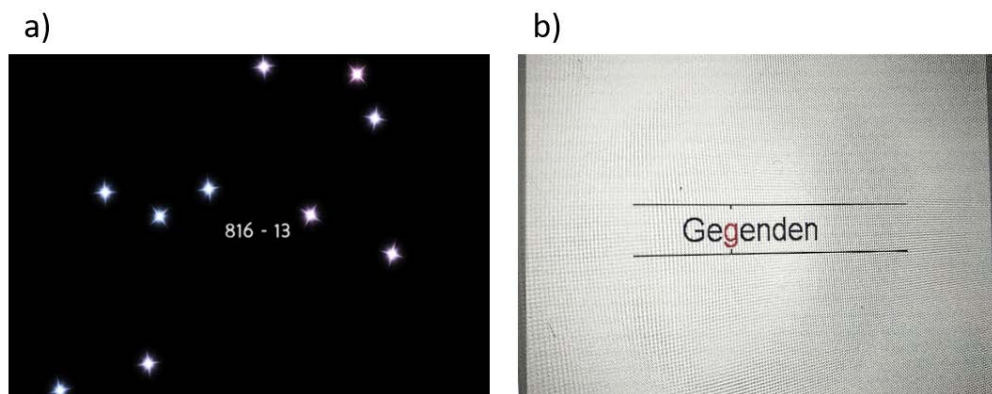


Fig. 4.1 Screenshots of the experimental paradigms. a) Mental workload paradigm: high workload trial with arithmetic subtraction assignment in the center of the screen, accompanied by visual distraction represented by colorful spots randomly moving around the screen. b) Text reading paradigm: Sequentially each word is presented in the RSVP reading interface (here ‘Gegenden’ meaning ‘areas’ in German). During reading the red highlighted letter is fixated. This letter is defined as the optimal point of recognition (ORP) for each word. The duration of presentation for each word is determined by the subjectively chosen presentation speed and factors such as word length or whether it is the end of a sentence.

4.3.3.1 Mental workload paradigm

The mental workload paradigm used for calibrating the task-independent predictive model was developed by Team PhyPA (Technical University Berlin). During this paradigm subjects alternate between two tasks to induce phases of low and high mental workload. In one task subjects focused on a crosshair in the middle of an otherwise black screen. In this condition subjects were instructed to relaxing their mind and keeping their eyes open. This task is intended to induce a low level of mental workload. To induce levels of high workload an arithmetic subtraction assignment appeared in the center of the screen. At its appearance subjects were instructed to silently subtract the number on the right side (range between 6 and 19, excluding 10 and 15) iteratively from the number on the left (range between 200 and 1200) until the assignment disappeared from the screen. In both tasks, 50 per-

cent of the trials were accompanied by visual distraction represented by colorful spots randomly moving around the screen. In previous studies, these distractions have proven to serve the purpose of dissolving dependencies of any class to ocular artefacts in the EEG resulting from visual differences in the tasks (Krol et al., 2016; Zhang et al., 2018). The rationale here is that ocular artefacts are controlled by the moving distractors to a high degree. In this way both classes (low and high workload) contain a similar, randomized amount of ocular artefacts such that the differences induced by such artefacts are not class dependent. Subjects completed 40 trials of tasks with low and 40 trials with tasks of high induced mental workload. Each trial had a length of 10 seconds.

4.3.3.2 Text reading paradigm

Each subject read twelve texts from the GEO/GEOLino Corpus (Hancke et al., 2012) for which we calculated scores of the two common readability formulas Flesch reading ease and the Flesch-Kincaid grade level. The corpus is a collection of 1066 texts in German from the magazine GEO, which has a focus on topics from nature, culture and science and the magazine GEOLino, which covers similar topics but is targeted at young people from age 8 to 14. Therefore, texts from GEO are generally more difficult to read than those from the GEOLino magazine. Six texts were chosen from each magazine, all covering similar topics about animals and their habitat. The overall average number of words per text was 493 (SD = 34.6 words). GEO texts had an average word count of 472 words (SD = 23.1 words) and GEOLino texts of 514 words (SD = 31.7 words). GEO texts had an average Flesch reading ease of 45.1 (SD = 2.4), which is equivalent to difficult texts on college level. The Flesch-Kincaid grade level of GEO texts was 10.9 (SD = .29). For GEOLino texts the average Flesch reading ease was 62 (SD = 1.38), corresponding to a readability suitable for 13 to 15 years old students. These texts had an average Flesch-Kincaid grade level of 7.9 (SD = .24).

All textual contents, except experimental instructions, were presented with the Spritz speed reading application interface. In the beginning of the paradigm subjects read a short text to familiarize themselves with text reading using the Spritz interface. Thereafter, subjects read short passages of a German novel and incrementally adjusted the text presentation speed to a level they felt comfortable reading at. The initial presentation speed was 200 wpm. After each passage subjects decided by button press whether the current presentation speed appeared too slow, too fast or just right. According to the subject's decision in the first presentation, speed was altered by 100 wpm in the direction favored by the subject (or stayed the same). For the next passages, the increment was decreased by half in each step

and again adjusted in the direction the subject wanted the presentation speed to alter. Presentation speed adjustment continued until the subject had decided that the current speed was comfortable for two successive text passages. The last speed value in wpm was set as the individually preferred presentation speed.

Subsequently each subject read all twelve texts. The six texts of each text difficulty category (easy vs. difficult) were randomly assigned to two groups of three texts. One of the two groups was presented in the individually adjusted presentation speed (normal category). The other group was presented at this speed raised by 40 percent to increase the level of mental workload (fast category). Beforehand, we assessed in a pre-test ($n = 6$, age 27.2 years ($SD = 3.8$ years), all native language German, five male) that an increase in presentation speed by 40 percent resulted in higher subjective workload ratings according to the Raw Task Load Index questionnaire (RTLX, Hart (2006)). This is a modified version of the NASA-TLX (Hart and Staveland, 1988), a standardized questionnaire to assess perceived mental workload on a Likert scale along six dimensions, lacking subjective weightings of the scales as used in the NASA-TLX. Between reading each block of three texts subjects could rest before starting the next block. The presentation order of the four text blocks was randomized. After each text block, subjects completed the RTLX questionnaire to assess their subjectively perceived workload. After each text, subjects answered three literal text comprehension questions. These questions were generated by taking facts stated within single sentences in the texts and turning them into questions (e.g. statement "With a diameter of 37 centimeters its eyes are the biggest in the animal kingdom." - generated question: "Which part of the giant squid is the largest of its kind in the animal kingdom?" - displayed answers: 1) teeth 2) nose 3) brain 4) eye). For each question, four possible answers were displayed only one of which was the correct choice. Subjects selected their answer by key press. In total each subject answered 36 literal text comprehension questions (3 questions x 12 texts).

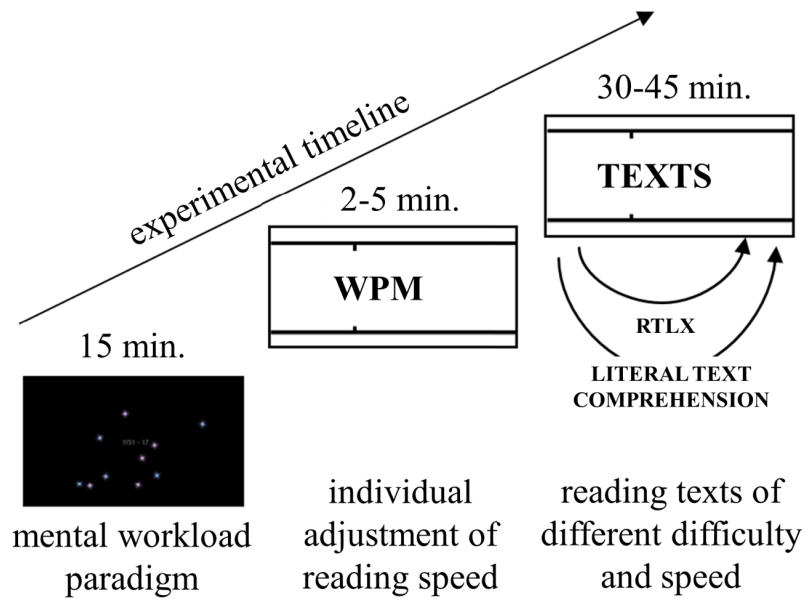


Fig. 4.2 Experimental procedure: Subjects first completed the mental workload paradigm. Afterwards they were familiarized with the speed reading presentation technique and adjusted the text presentation speed to a level of words per minute at which they felt comfortable reading. Then in the text reading paradigm subjects read all twelve texts in blocks of three. Block order was randomized and the three texts in a block were either easy or difficult texts presented at either the speed adjusted by the subject or with an increase of 40 percent in speed. After each text subjects answered three literal text comprehension questions. Following each block subjects completed the RTLX questionnaire.

4.4 Data Analysis

4.4.1 EEG data

4.4.1.1 Calibration of the task-independent predictive model

To calibrate a subject-dependent passive BCI classifier for the task-independent predictive model, features from the spectral domain of the individual EEG were extracted using filter bank common spatial patterns (fbCSP, Ang et al. (2008)). Neuroscientific findings provide evidence that frequency bands from 4-7 (theta band) in fronto-central areas and 8-13 Hz (alpha band) in parietal areas of the human scalp show differences with increasing mental workload, in general (see Gerjets et al. (2014); Krause et al. (2000); Pesonen et al. (2007); Scharinger et al. (2015, 2017)). Typically, increased mental workload is associated with an increase in frontal theta and a decrease in parietal alpha activity (Gevins and Smith, 2000; Gevins et al., 1997).

Along these findings, each of both frequency bands were selected as a filter bank for fbCSP. In each trial, for each electrode, the EEG data was preprocessed by filter-

ing the band power in each of the two frequency bands, serving as input for fbCSP. This approach resulted in a set of *spatial filters* to optimally extract features to discriminate between low and high workload. Each set contained 2x3 spatial filters per frequency band, where each subset of 3 filters was optimized for extracting maximal bandpower in one class while keeping the bandpower minimal in the other class, reflecting the above-mentioned concept of spectral indicators for mental workload. EEG data projected through each of the 12 resulting filters served as features for classification.

Data from each trial of 10 seconds was cut into consecutive epochs of one second length, defining the *training trials* for classifier calibration. Each training trial served as an instance for either high or low workload, according to the class of whole trial. For each epoch, 3 features were extracted with fbCSP for each class and each frequency band, as mentioned above, resulting in a set of (3x2x2 =)12 dimensional feature vectors for each training trial. On the resulting set of 12 dimensional features for each class, regularized linear discriminant analysis (LDA) was trained as a classifier that discriminates between high and low workload. As we expect this model to be task independent, the accuracy of the classification on the calibration data provides an estimate for the accuracy of the calibration on the text reading data. Therefore, a means of estimating classification accuracy was derived by performing a (5,5) *cross validation* on the calibration data. The data was randomly divided into 5 partitions (indicated by the second number in (5,5)) from which each was considered to be test data while the remaining four partitions served as calibration data for the classifier. The classifier was then trained on the calibration data and tested on the remaining partition, resulting in a classification error for this run of the cross validation. The average of all runs of the cross validation indicated the estimate of the classification error. This whole process was repeated 5 times (the first number in (5,5)) with different random selections of the partitions to minimize the error resulting from bias in the randomization. The average of each full cross validation served as a final estimate of the classification error expected for the application of the classifier on the text reading data. With this, the signal processing and machine learning for the task-independent predictive model is defined.

As described in section 2.5.4 these filters are optimal for class discrimination, they cannot be interpreted spatially from a neuroscientific perspective. By transforming them into so-called *spatial patterns*, information about scalp projections of cortical activity that is relevant for classification can be derived and interpreted neuroscientifically (Haufe et al., 2014). For each feature a neuroscientifically interpretable spatial pattern can be generated, identifying the electrode activity that was relevant to discriminate the classes in the according frequency band. In that way, information used by the classifier to discriminate between classes can be identified

and used to validate the classification approach. Spatial patterns generated in this process were averaged over the three patterns in each subset and then over all subjects resulting in four grand average patterns displaying the topography for the two classes of low and high workload in the alpha and the theta frequency band.

4.4.1.2 Classifier application to text reading data

For all subjects with an estimated classification accuracy of 60 percent or higher the calibration process was deemed as being acceptable as it was statistically better than random (Müller-Putz et al., 2008). We used BCILAB's built in function *onl_simulate* to apply the individually calibrated task-independent predictive model discriminating mental workload to raw data of all twelve texts per subject. The model was applied to epochs of 1 second length, serving as test trials, starting upon the onset of each displayed word of a text. The volume of the classes ranged between 1347 and 1635 samples, as each class was populated with samples from three texts and the number of words per text ranged between 449 and 545 (see second part of section Stimuli and procedure 4.3.3), according to the number of words in each text. The output of that classification was then annotated to the according word, indicating the workload that was predicted by the model at this time. Epochs of the test trials overlapped, as there was more than one word displayed within one second during the reading task. However, deriving predictive values from overlapping epochs is unproblematic in our case. We are not performing binary classification but generating a continuous development of predictive value over time – and are not aiming at classifying single words, but rather text sections consisting of multiple words with respective workload predictive values. The resulting predictive values for each word, ranged between 0 (low predicted workload) and 1 (high predicted workload) This procedure resulted in a prediction of the mental workload for each word in each text.

Subject specific text-group averages of predictive values were compared to the intrinsic difficulty of each text by means of a two-way, repeated measures ANOVA. The two within factors were text difficulty (easy vs. difficult) and presentation speed (normal vs. fast).

To obtain further insights into possible effects in the repeated measures ANOVA detailed above, predictive values for each text category generated from each individual task-independent predictive model were subjected to permutation tests (see section 3.2.1.2 with 5000 permutations. This number of permutations was chosen to ensure a precision of the p value better than 0.001 in a confidence interval of 99 percent, according to Ojala and Garriga (2010). Predictive values from all subjects according to text difficulty (easy vs. difficult) and presentation speed

(normal vs. fast) were tested against those from all other categories, resulting in $6 (= \binom{4}{2} = 3 \times 2 \times 1)$ permutation tests. Five of these tests were one-tailed because we assumed that easy texts would result in lower predictive values than difficult texts. Also, predictive values from texts of comparable difficulty were expected to be lower when presented at normal speed than predictive values of texts presented at an increased speed. Easy texts presented at normal speed were expected to have lower predictive values than those of difficult, fast presented texts. As no assumption regarding mental workload could be made, the permutation test comparing predictive values of easy/fast-speed texts and difficult/normal-speed texts was two-tailed.

Finally, we analyzed the predictive values on single-subject level by performing further one-tailed permutation tests. We hypothesized that predictive values for easy texts would be lower than those of difficult texts when read at the same presentation speed. Further we expected texts of the same text difficulty to have lower workload predictive values for texts read at normal speed than at a faster pace. Finally, we expected easy texts read at normal speed to have lower predictive values than difficult texts read at fast speed. The five resulting permutation tests per subject were computed with again 5000 permutations using all predictive values from the three texts belonging to the same text category.

To obtain further insight into classification performance on a single-subject level we performed an analysis of how to determine the amount of read words that need to be classified to make a reliable discrimination between text categories. For each word position we averaged the predictive values derived from texts of the same category (three for each text difficulty or speed) at this position to correct for the intra-text variance of word and text complexity at that position in that category. This allows us to calculate the *cumulative predictive value* for each word position in each category by averaging these corrected predictive values up to the selected word position: For each word w_n^k in text T_k , that is from a certain text category c , the cumulative predictive value $c p$ at position n is derived by

$$c p_n^c = \sum_{l=1}^n \sum_{k=1}^3 \frac{c l a(w_l^k)}{3n}$$

Where $c l a(w_l^k)$ denotes the predictive value derived for the word w_l^k by linear discriminant analysis. Let E be the segment of EEG data recorded for word w_l^k and P be the preprocessing of that data as described in the section 4.4.1.1. Then the application of filter bank CSP on the preprocessed data results in a 12-dimensional feature vector $\text{fbCSP}(P(E))$, as described in section 4.4.1.1. Applying LDA to this feature vector then results in the single digit predictive value $c l a(w_l^k) = \text{LDA}(\text{fbCSP}(P(E)))$.

We plotted the development of the cumulative predictive values $(c p_n^c)_{n \in \{1, \dots, 422\}}$ for each word position in each text category. As the shortest text contained 422

words we only considered the cumulative values of each text category up to the 422nd word. We plotted all four combinations of factor comparisons (difficulty (easy vs. difficult) vs. presentation speed (normal vs. fast)) for each subject. Each trained task-independent predictive model is a probabilistic operator with a certain error rate. Hence, each (cp_n^c) should only be considered as an estimate of the actual workload at that word position. We calculated the individual variance of the error of the predictive values for each subject by applying the mental workload classifier again to the calibration paradigm data. The standard deviations of the predictive values for each class were averaged and served as a criteria for the error range for the cumulative predictive values $(cp_n^c)_{n \in \{1, \dots, 422\}}$: For each subject the *classifier deviation* is defined as $D := \frac{std(C^1) + std(C^2)}{2}$, where (C^k) denotes the standard deviation of the predictive values for class k . Since it is taking all predictive values considered into account, cp_{422}^c can be seen as the most informed estimate of the workload associated to that text category. The *stability word position* W is defined as the highest word position where the cumulative predictive value exceeds the (rather conservative) criteria of one quarter of the classifier deviation from cp_{422}^c : For each subject and text category c , $W^c := cp_M^c$, with $M = \max_n (|cp_n^c - cp_{422}^c| > \frac{D}{4})$. The cumulative predictive values for each word position following W^c can then be considered to be an estimate for the workload associated to the corresponding text that is as good as that from cp_{422}^c . In other words, the text position of the stability word serves as an estimate of how many words of a text a single person must read such that the task-independent predictive model can provide a best estimate for mental workload induced by that particular text category. For each subject, we plotted the highest text position from the four determined stability word positions (for the four different text categories) as the general stability word position. We assume that after this text position, we can use the cumulative value at this text position as a reliable indicator of induced mental workload for all text categories.

4.4.2 RTLX

Subjective workload ratings assessed in the RTLX questionnaire were converted to workload scores according to NASA-TLX procedures (Hart, 2006). These scores were then subjected to a two-way repeated measures ANOVA. The two within factors are text difficulty (easy vs. difficult) and presentation speed (normal vs. fast).

4.4.3 Literal text comprehension

For each subject we collected the number of incorrect answers (total of 9 questions per block, three questions per text) per text category. We then computed a two-way repeated measures ANOVA, with the two within factors text difficulty (easy vs. diffi-

cult) and presentation speed (normal vs. fast) for the counts of correctly answered questions.

4.5 Results

4.5.1 EEG data

4.5.1.1 Mental workload paradigm

Table 4.1 gives an overview of classification error rates estimated from cross validation in the mental workload paradigm of all initial thirteen subjects. The average estimated classification error rate was 25.9 percent (SD = 4.9). We performed a check to see whether individual results provide evidence that the task-independent predictive model had classification accuracies that are significantly better than random classification, taking the confidence interval into account (see Müller-Putz et al. (2008) for a detailed description of this approach). The results show that, for all but one subject, classifications are expected to be clearly better than random classification. Therefore, this subject was excluded from further analysis.

Table 4.1 Estimated classification results of the mental workload paradigm for all initial thirteen subjects. Obtained error rates (ER) from (5,5) cross validation in percent and their standard deviations (SD), resulting from variance in the classification error estimated from each run of the cross validation are reported.

subject	calibration error	SD
1	23.3	2.0
2	31.0	12.0
3	13.9	3.1
4	25.0	3.4
5	27.1	3.2
6	28.6	6.5
7	10.9	5.5
8	27.8	3.2
9	31.8	2.4
10	27.0	8.4
11	35.8	5.7
12	13.3	4.2
13	41.8	4.1
average	25.9	4.9

Figure 4.3 depicts that the spatial patterns resulting from the classifier model show a clear focus on fronto-central electrodes for the theta band and on parietal electrodes for the alpha band, providing evidence that the filters derived by fbCSP

are in accordance with the underlying neurophysiology.

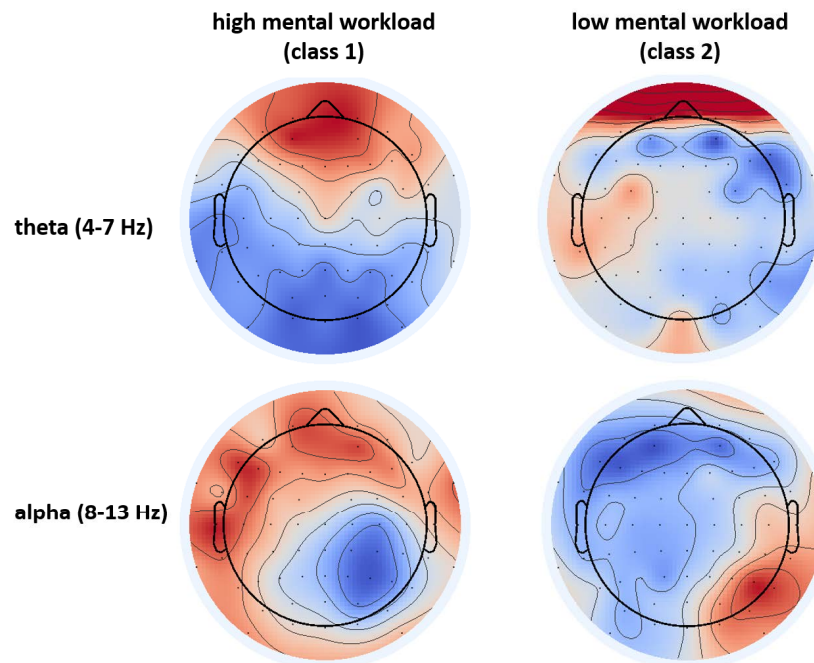


Fig. 4.3 Spatial patterns available to the mental workload task-independent predictive model averaged over all twelve subjects where classification was discriminative. For the high mental workload condition (class 1) the patterns show a clear increase in frontal theta and a decrease in (lateral) parietal alpha. These results are in accordance with findings from scientific literature regarding the effects of increased mental workload.

4.5.1.2 Text reading paradigm

Subject specific text-group averages of predictive values are displayed in figure 4.4. The two-way repeated measures ANOVA (see figure 4.6 a)) performed on mental workload predictive values yielded significance for the main factor text difficulty, $F(1,11) = 5.04$, $p = .046$. Predictive values of difficult texts ($M = .618$, $SD = .272$) were higher than predictive values for easy texts ($M = .553$, $SD = .259$). The main factor text presentation speed was significant as well, $F(1,11) = 11.64$, $p = .006$. Predictive values of texts presented at normal speed were lower ($M = .559$, $SD = .267$) than those of texts presented at the increased speed ($M = .616$, $SD = .264$). There was no significant interaction, $F(1,11) = .036$, $p = .854$.

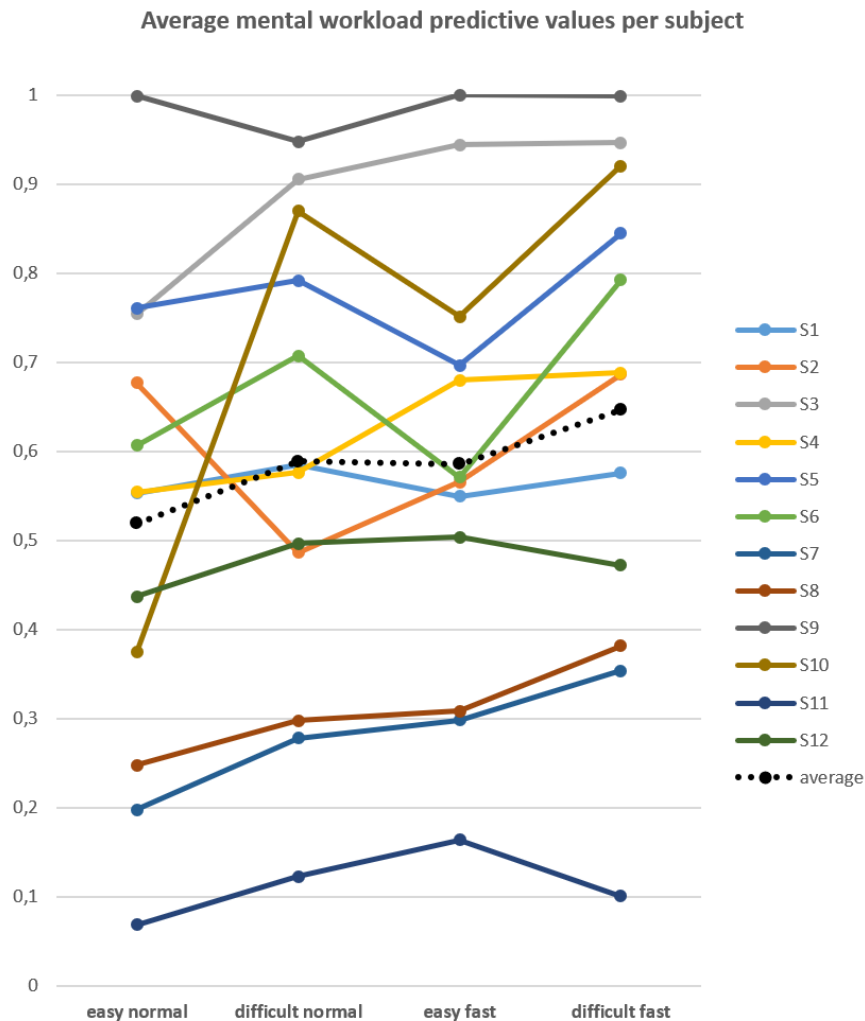


Fig. 4.4 Subject-wise predictive values of mental workload derived from application of the task-independent predictive model trained on data from the mental workload paradigm to text reading data. For each text reading condition all predictive values of all three texts belonging to the respective condition were averaged. The dotted line displays the grand average of all subjects ($n = 12$). The results of the ANOVA and the related permutation tests across subjects are validating the general hypotheses (as depicted in the dotted line), while according significant differences between conditions within subjects (solid lines) were also found in subject-specific permutation tests.

The results of five from the six permutation tests across subjects were highly significant (all p s $< .0001$). Only in the case where predictive values of easy texts presented at fast speed tested against predictive values from difficult texts presented at normal speed, results were not significant ($p = .340$). However, absolute values of observed differences in predictive values between text conditions ($M = .058$, $SD = .036$) were much smaller in all tests than variances of predictive values within text conditions ($M = .265$, $SD = .011$). Effect sizes were accordingly small ($M = .181$, SD

= .069), ranging between .129 (easy fast vs. difficult fast) and .301 (easy normal vs. difficult fast) in tests with significant results.

VP	permutation test result	easy/normal vs difficult/normal	easy/fast vs difficult/fast	difficult/normal vs difficult/fast	easy/normal vs easy/fast	easy/normal vs difficult/fast	stability word position	workload paradigm classification: error rate (SD) in percent
1	p-value	< 0.005	< 0.01	> 0.01	> 0.01	> 0.01	103	22.3 (2.0)
	observed difference	-0.0317	-0.0263	0.0086	0.0032	-0.0230		
	effect size	0.1040	0.0892	0.0289	0.0108	0.0778		
2	p-value	> 0.01	< 0.0005	< 0.0005	> 0.01	> 0.01	94	31.0 (12)
	observed difference	0.1904	-0.1209	-0.2004	0.1109	-0.0010		
	effect size	0.5113	0.3345	0.5445	0.3032	0.0285		
3	p-value	< 0.0005	> 0.01	< 0.0005	< 0.0005	< 0.0005	85	13.9 (3.1)
	observed difference	-0.1513	-0.0019	-0.0407	-0.1901	-0.1920		
	effect size	0.7782	0.0163	0.2854	1.1114	1.1178		
4	p-value	> 0.01	> 0.01	< 0.0005	< 0.0005	< 0.0005	159	25 (3.4)
	observed difference	-0.0216	-0.0086	-0.1122	-0.1253	-0.1338		
	effect size	0.0666	0.0281	0.3556	0.3989	0.4218		
5	p-value	< 0.005	< 0.0005	< 0.0005	> 0.01	< 0.0005	157	27.1 (3.2)
	observed difference	-0.0308	-0.1484	-0.0530	0.0645	-0.0839		
	effect size	0.1059	0.5330	0.1100	0.2121	0.3103		
6	p-value	< 0.0005	< 0.0005	< 0.0005	> 0.01	< 0.0005	199	28.6 (6.5)
	observed difference	-0.1005	-0.2209	-0.0851	0.0353	-0.1856		
	effect size	0.3661	0.8609	0.3438	0.1245	0.7081		
7	p-value	< 0.0005	< 0.0005	< 0.0005	< 0.0005	< 0.0005	106	10.9 (5.5)
	observed difference	-0.0805	-0.0553	-0.0752	-0.1003	-0.1557		
	effect size	0.3018	0.1902	0.2655	0.3654	0.5733		
8	p-value	< 0.0005	< 0.0005	< 0.0005	< 0.0005	< 0.0005	106	27.8 (3.2)
	observed difference	-0.0498	-0.0730	-0.0839	-0.0607	-0.1337		
	effect size	0.1554	0.2152	0.2502	0.1870	0.4068		
9	p-value	> 0.01	> 0.01	< 0.0005	< 0.001	> 0.01	7	31.8 (2.4)
	observed difference	0.0512	0.0013	-0.0512	-0.0013	0.0001		
	effect size	0.5696	0.0969	0.5471	0.1264	0.0023		
10	p-value	< 0.0005	< 0.0005	< 0.0005	< 0.0005	< 0.0005	300	27.0 (8.4)
	observed difference	-0.4952	-0.1691	-0.0505	-0.3765	-0.5457		
	effect size	1.4862	0.5934	0.2212	0.9657	1.7869		
11	p-value	< 0.0005	> 0.01	> 0.01	< 0.0005	< 0.0005	163	35.8 (5.7)
	observed difference	-0.0536	0.0632	0.0220	-0.0948	-0.0316		
	effect size	0.3161	0.3301	0.1217	0.5256	0.1979		
12	p-value	< 0.0005	> 0.01	> 0.01	< 0.0005	< 0.005	91	13.3 (4.2)
	observed difference	-0.0592	0.0314	0.0241	-0.0665	-0.0351		
	effect size	0.1845	0.0963	0.0730	0.2099	0.1071		
		9 of 12	7 of 12	9 of 12	8 of 12	9 of 12		

Fig. 4.5 Results of subject-wise one-tailed permutation tests performed on predictive values. For each test the category first named in the table was expected to have lower workload predictive values. We also noted the individual stability word position for predictive values (indicating how many words need to be read to make a reliable estimation of the average predictive value the reading of a text induces in the specific subject), as well as the classification error rate of the mental workload classifiers. The tests were Bonferroni corrected for multiple testings in each subject, resulting in a significance level of .01

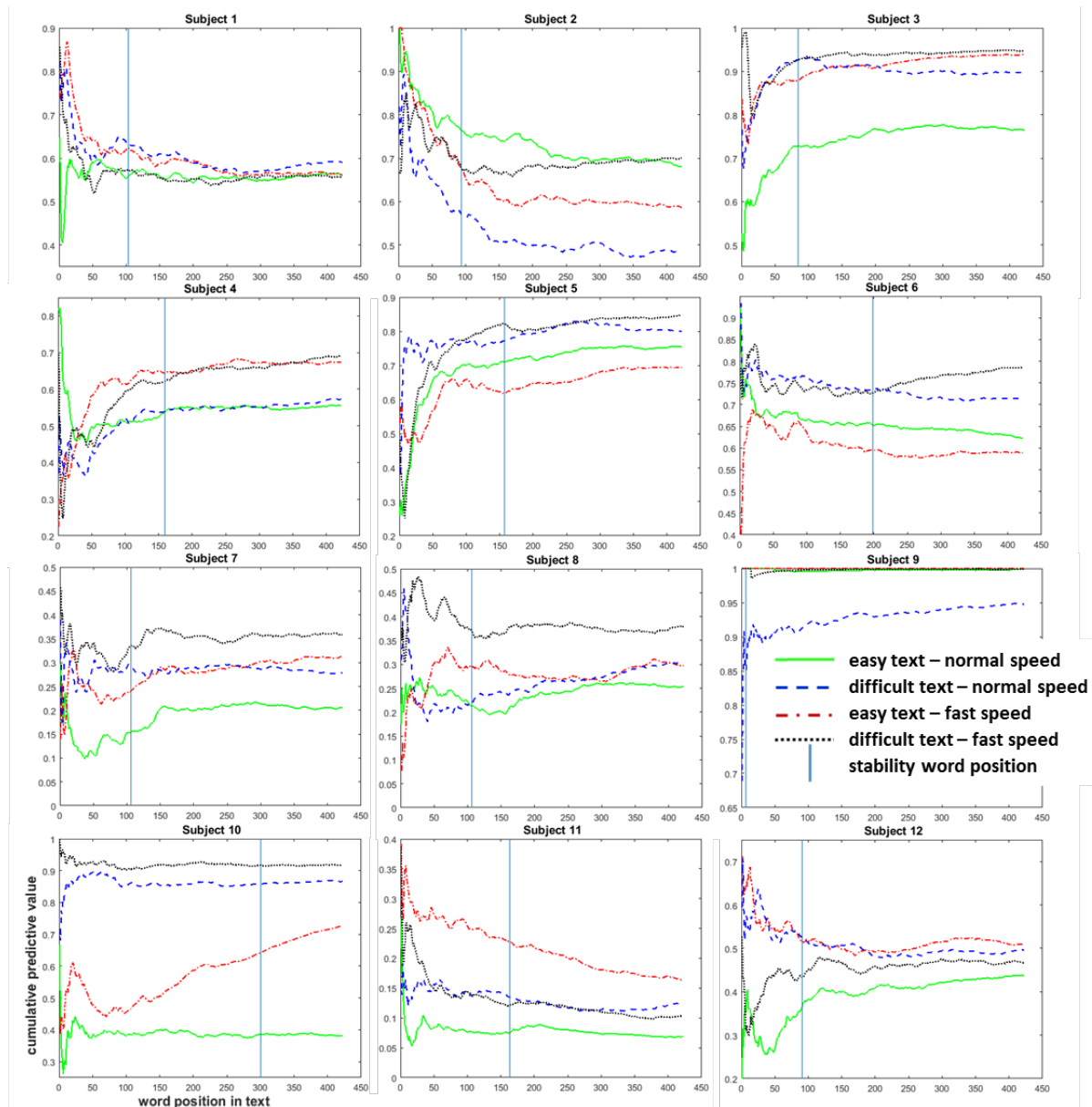


Fig. 4.6 Cumulative predictive values derived from the predictive models of the mental workload classifiers. Predictive values of all three texts per category were averaged and cumulated for each successive word in the texts. The development of cumulative predictive values for all four text reading conditions is shown for the first 422 words in the texts. On the x-axis the text position of the word is displayed. On the y-axis the averaged cumulative predictive values are noted. The individual stability word position (highest of all for text categories) is indicated by a vertical black line.

On single-subject level, one-tailed permutation tests (see table 4.5) showed significant results for 9 of 12 subjects when predictive values of easy texts and difficult texts, both read in normal speed, were tested against each other. 9 tests also yielded significance when easy texts read at normal speed and difficult texts read fast were compared. When we compared predictive values from difficult texts read at nor-

mal and fast speed, 9 of 12 permutation tests were significant. For easy texts read in normal and fast speed 8 of 12 tests yielded significance. The lowest count of significant permutation tests was reached when we tested predictive values of easy and difficult texts read at both fast speed against each other – here only 7 of 12 tests had significant results. The average number of significant permutation test results was 3.67 (SD = 1.07) per subject out of 5 tests. Overall effect sizes of significant permutation test results are close to medium with an average effect size of $.444$ (SD = $.379$). The individually determined stability word position (see figure 4.6) was at text position 130.83 (SD = 72.72) on average.

4.5.2 RTLX

The two-way repeated measures ANOVA (see figure 4.7 b)) performed on subjective workload ratings revealed significance for the main factor text difficulty, $F(1,11) = 6.98$, $p = .023$. Overall difficult texts ($M = 24.98$, $SD = 10.76$) received higher mental workload ratings than easy texts ($M = 22.78$, $SD = 10.51$). Results for the main factor presentation speed were significant as well, $F(1,11) = 6.48$, $p = .027$. Texts presented at the normal ($M = 21.95$, $SD = 10.18$) self-adjusted reading speed received lower RTLX ratings than texts presented with an increase of 40 percent ($M = 25.81$, $SD = 11.09$). The interaction was not significant, $F(1,11) = 1.87$, $p = .199$.

4.5.3 Literal text comprehension

Individually adjusted text presentation rates from all subjects ranged between 150 and 500 wpm. The average adjusted reading speed was 269 wpm (SD = 109 wpm). The ANOVA performed on incorrect answers given to literal text comprehension questions (see figure 4.7 c)) revealed neither significant main effects, nor an interaction effect, all $ps > .249$. On average subjects answered 2.6 (SD = 1.7) questions out of nine per text block incorrectly. An average of 2 (SD = 1.8) incorrect answers was given for easy texts and 3.1 (SD = 1.3) for difficult texts presented at normal speed. For texts blocks with an increased presentation speed, questions on easy texts were answered 2.7 (SD = 1.5) times incorrectly on average and difficult texts 2.8 (SD = 1.9) times.

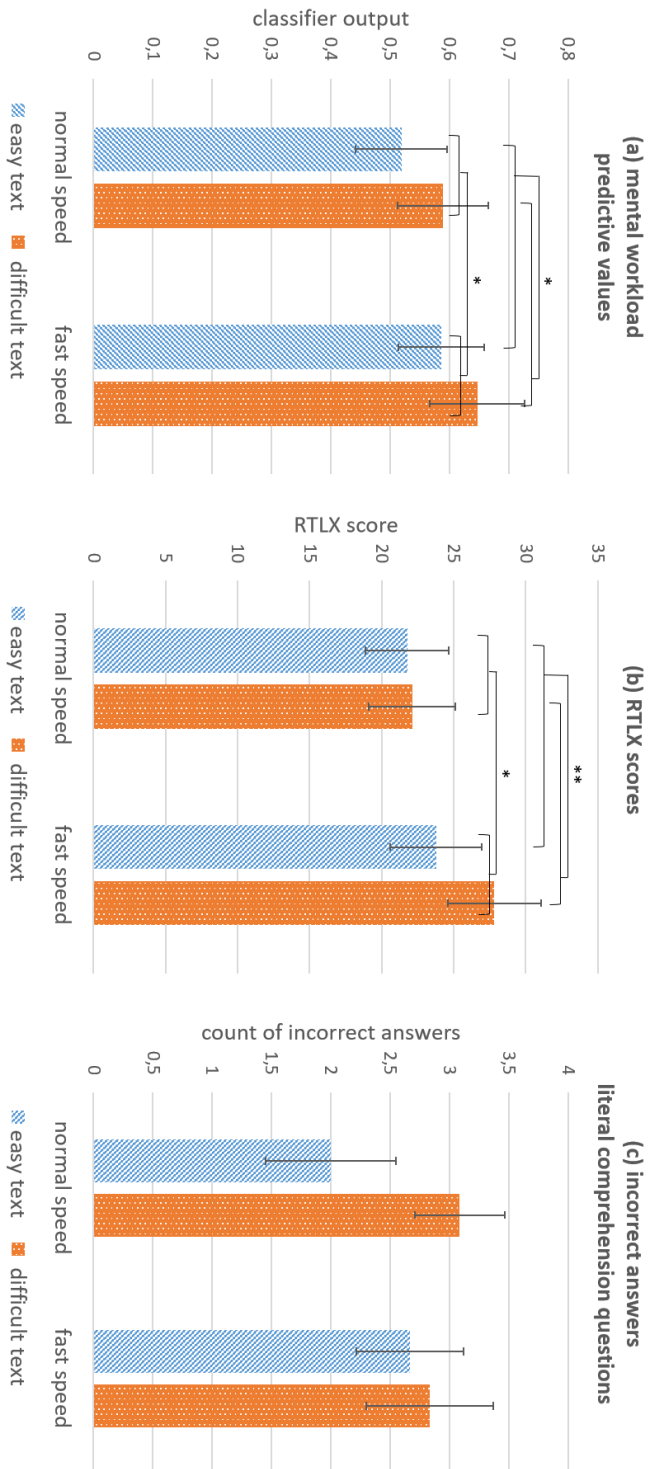


Fig. 4.7 (a) Predictive value averages (n= 12) of the task-independent predictive model applied to EEG data of each word of the twelve texts subjects read, (b) averages (n= 12) of RTLX ratings for the different text difficulties and text presentation speeds and (c) averaged (n= 12) numbers of incorrectly answered literal text comprehension questions. Error bars represent standard deviations. Significant results of the ANOVAs are indicated.

4.6 Discussion

The aim of this investigation was to examine whether a task-independent predictive model for mental workload detection shows a sensitivity high enough to identify the individual difficulty induced by reading a specific text. The results of the presented experiment show that an individually calibrated model shows significant differences between reading easy and difficult texts at different presentation speeds.

Regarding text readability, results from EEG data suggest that we can significantly distinguish easy from difficult texts, even on single-subject level, even though variances of mental workload predictive values within and between subjects are high. High variances between subjects most probably are a result of different intra-subjective ratios between arithmetic and text reading abilities.

To our knowledge no other study has yet investigated whether text difficulty and reading speed can be directly deduced from mental workload assessed by EEG. Studies (e.g. Schultheis and Jameson (2004)) have investigated changes in EEG activity induced by reading. They provided a descriptive analysis of that brain activity, but did neither identify sensitive and specific cognitive processes – like mental workload – related to that activity, nor did they define and investigate a measure from which text difficulty or reading speed could be deduced.

The neuroadaptive measure defined here showed overall that, predictive values were higher for words from difficult texts than for words from easy texts. Moreover, on single-subject level for nine out of twelve subjects, mental workload predictive values were significantly higher for difficult texts than for easy texts when reading at a self-adjusted speed level. If difficult texts were read with an increased speed, predictive values for mental workload were even higher and more distinct from the lower predictive values of easy texts read at normal speed. The latter was the case for nine out of twelve subjects as well. When we compared predictive values of easy and difficult texts both read at an increased speed, only seven out of twelve permutation tests yielded significant results.

Regarding differences in mental workload induced by different text presentation speeds we found a highly significant effect for increased presentation speeds on mental workload predictive values. On single-subject level, in nine out of twelve cases mental workload predictive values were higher for difficult texts when they were read at an increased speed compared to reading at a self-adjusted speed level. For easy texts, in eight out of twelve subjects' predictive values were considerably higher when reading at increased texts presentation speed opposed to reading at a self-adjusted speed. As the chosen texts were written for specific target groups (educated adults and young teenagers), it can be assumed that the difficulty of each

category is aimed at a certain educational level. Furthermore, measures for text complexity confirmed a difference between easy and difficult texts (see section Materials and Methods – Stimuli and procedure 4.3.3– 2. Text reading paradigm) and an increase of reading speed should add a stable amount of mental workload for the reader. Taken together, it can be assumed that each full text induces a certain level of mental workload when read, and that these levels can be identified by the used predictive models. The classifier outputs are in line with this assumption as shown in figure 5 and table 2. The curves in figure 5 stabilize after about 130 words, showing that the classifier outputs converge to a clear estimate of the induced mental workload, already after about 30 seconds of reading. In addition, a clear, expected difference in measured workload can be seen across subjects between the different text categories, as validated by the ANOVA and depicted in figure 4. In conclusion, the results of this study are fully in line with the expected effect of the stimuli on the mental workload of the readers, providing a high validity to the used task-independent approach based on passive BCI.

Results of the ratings given in the RTLX questionnaire assert that readers perceive a rise in reading speed of 40 percent as an increase in subjectively experienced mental workload. Also, reading difficult texts was associated with higher scores for mental workload as compared to easier texts. We therefore conclude that the subjective experience of mental workload was in accordance with the properties of the predictive values we obtained from the predictive model for each subject with regard to changes in text readability and reading speed.

Since no statistical differences in literal comprehension were shown between different text difficulties and presentation speeds, it can be concluded that an increased presentation speed did not lead to lower levels of literal comprehension in our study. On average, two thirds of the comprehension questions within each text block were answered correctly by subjects. Thus, the results from literal comprehension questions indicate that subjects read all variations of texts attentively and with the same level of basic understanding. However, as we did not include more complex comprehension questions that require for elaborated inferences, we cannot rule out that higher comprehension processes might have suffered from increased reading speeds.

The self-selected presentation speed for the normal presentation showed a large variance between subjects. The aim of selecting the individual speed was to generate a speed that is comfortable for each individual reader, neither overloading nor boring them. The results of the text comprehension analysis showed a high comprehension rate with low variance, indicating that subjects could follow the texts easily. This leads to the conclusion that the presentation speeds were indeed comfortable for the subjects, even though the presentation speeds differed strongly be-

tween readers. The overall duration of our experiment was a little less than an hour on average. We took several counter measures to avoid fatigue overall, as well as possible effects of it. Firstly, participants were allowed several breaks throughout the experiment to recover from tasks they were performing. Even if fatigue or high workload were not avoided by those breaks, they should not have an effect on the results we saw from the experiment, as the order the four text categories presented to subjects was randomized and the overall duration of the experiment was about the same for all subjects, so the induced fatigue should not vary among subjects.

Altogether, the results showed that the task-independent predictive models are capable of identifying different levels of text readability and presentation speed on single-subject level. We proposed an estimate of how many words need to be read to make a reliable estimation of the average mental workload associated to a text. This criterion suggests that an average count of about 130 words would be sufficient.

In conclusion, the presented results allow to add text readability and presentation speed to the task domains the task-independent mental workload classifier introduced in Krol et al. (2016) and further investigated by Zhang et al. (2018) can be successfully applied to.

The presented approach deviates from often-used practices of binary classification based on BCI and aims at the identification of the levels of workload on a continuous scale. Results presented here support the idea that, in future applications of this approach, the subjective reading difficulty of a certain section in a text can be identified based on estimations of the induced mental workload. It, hence, is not necessary to define certain categories of difficulty for texts, as done in the presented experiment. Instead, the subjective, momentary reading difficulty of a text can be traced over time on a continuous scale.

However, it remains unclear how well the predictive model would perform for texts whose readability is more similar than that of the texts used in this study. The precision of the mental workload classifier should be further investigated in future research by presenting subjects with a larger spectrum of readability levels. Furthermore, smaller steps of presentation speed alternation should be examined in order to investigate how mental workload predictive values are changed at presentation speeds lower than an individually adjusted level. It would also be of interest to determine whether or not predictive values behave linearly with changes in text readability and presentation speed.

In a real-world reading scenario, the main goal of the user is to read a specific text. From a user experience point of view, it would be bothersome to have to invest additional time and effort to enable a more individualized user support for the main task of reading. For the implementation of a neuroadaptive reading interface

currently such additional effort must be invested by the user, as he undergoes an initial calibration phase to collect training data for the task-independent mental workload classifier. In the future, this time consuming step in creating a neuroadaptive reading environment could become unnecessary through the establishment of a mental workload classifier that is not only task-, but subject- independent as well. In such a universal classification (Zander (2012), p. 96ff. and p. 142ff.) a classifier can be defined that can be transferred between sessions and even between subjects, leading to reduced calibration time for an individual subject or even making it obsolete. Approaches towards such universal classification focusing specifically on mental workload are currently investigated (see e.g. Durkee et al. (2016); Zander (2012)). Another approach for reduction of calibration time is transfer learning between calibration sessions (Jayaram et al., 2016) which has already been applied to motor imagery classification contexts (J. M. Krauledat (2008), p. 90ff.). In future investigations such approaches for a reduction of calibration time should be investigated for the here presented neuroadaptive reading environment.

In addition, it would be of interest how well the results from this study can be generalized to other forms of reading and text presentation techniques (as e.g. traditional continuous text reading). We assume that in further studies other reading scenarios could be investigated by additional collection of eye-tracking data. This would enable us to map which word of a text currently is read by a subject to the corresponding EEG data we record.

Though exceeding the scope of this study, it should be further investigated in what way mental workload is affected by RSVP reading itself. It has already been found that workload can be increased when long text passages are read with RSVP presentation (Öquist and Goldstein, 2003) and mental workload in general is higher for RSVP reading compared to traditional (continuous) reading (Gannon et al., 2016). An analysis and comparison of RSVP reading and traditional reading regarding mental workload is not a part of our work presented here as it is outside of the realm of the research questions we pose, which is restricted to RSVP reading only. In a follow-up study, where the generalizability of our results to traditional reading will be tested, this comparison of reading techniques would be investigated in more depth.

Moreover, in this feasibility study the subject's reading fluency was not assessed. Hence, possible effects of reading fluency on the classifier performance and workload prediction was not analyzed. In a follow-up study with more subjects the effects of reading fluency should be investigated.

While effect sizes on individual data are at a good level, those between subjects are rather small. This can be seen as a result of the high variance of the predictive values between subjects, as shown in figure 4. Currently, the reason for this high

variance is unclear. We assume, though, that it is a result of differences in mathematical abilities and reading skills, which should be investigated in future studies.

As reproducibility of results is relevant, specifically for new approaches as presented here, all preprocessed data and analyses code will be linked at the website of the Society for Neuroadaptive Technology (www.neuroadaptive.org).

4.7 Conclusion

Neuroadaptive systems have the potential, if well designed, to decrease mental workload for users during interaction with a system (Byrne and Parasuraman, 1996). Such a system can continuously learn while the user interacts with the machine and thereby build and refine a user model that allows for individualization of software and contents over time. Such a model would reflect highly subjective aspects of information that are not typically included in methods relying on large amounts of objective data. Statistical inference (see Vapnik (1999b)) based on large amounts of data can be used to calculate general readability metrics, but leave no room for adaptations based on individual, momentary situational changes (such as changes in a person's mindset). It is precisely this kind of transient information, however, that is paramount for correct situational interpretation in general (Hudlicka and Mcneese, 2002). We propose that the approach presented here could be used to complement models built through statistical inference on large amounts of data by providing information about the subjective, momentary interpretation of the human. Technology could then not only learn to understand general patterns that exist in our world, but also gain a better understanding of the momentary mindset of individual human beings in this world. The study presented here provides an additional, supportive argument – with the example of mental workload – for this vision.

We demonstrate that predictions made by the task-independent predictive model trained to identify levels of workload can be utilized in order to assess differences in text readability on a single-subject level. Additionally, changes in mental workload induced by different levels of text presentation speeds are also reflected in the output of the model. These levels can be distinguished through the classifier's predictive values of about the first 130 words read, which took about 30 seconds in our experiments. With that, the system can gain an understanding of the current effect a text has towards the reader's cognition and adapt itself accordingly.

Even though questions regarding the precision of the mental workload classifier in practical contexts remain to be answered, the results of our study suggest possible benefits this measure could bring in the form of a neuroadaptive system for (speed) reading applications.

A user model with a classifier trained over time to identify text readability levels for an individual user could automatically adapt parameters of the reading application in real-time, e.g. the presentation speed. Conducting such adaptations within a closed-loop system between reading application and reader can elicit changes in the user's mental state. These changes can then be fed back to the system again to perform further adaptation or to learn from the induced changes in mental user

state for future adaptation. The implementation of a neuroadaptive reading application could enable more effective, individualized reading at a balanced level of workload. Readers should neither be over- nor underloaded while reading for an optimal state of mind to process and store the information read. This was firstly discussed in instructional design theories (e.g. Vygotsky and Cole (1978); Salomon (1984)). The main problem for the application of these theories is to measure the current, actual mental workload, which was difficult so far. The idea of using neuroadaptive interfaces to solve this problem is discussed for the first time in a general sense in Gerjets et al. (2014). The work presented here builds up on these concepts and describes an approach that allows for the assessment of mental workload while reading a text. The next steps will be the use of this information in an adaptive reading interface, closing the loop and keeping the reader in an improved cognitive state, which can be done in different ways. Apart from presentation speed changes, possible adaptation of readability levels might include real-time exchange of words or restructuration of the text material. If the system has e.g. detected that the current material is difficult for the user to read, infrequent words could be automatically exchanged by easier synonyms. Furthermore, sentences could be restructured in order to reduce sentence length. Further possibilities for real-time text simplification have been discussed in Crossley et al. (2011).

This study shows a new example of how passive Brain-Computer Interfaces can be used to develop innovative Neuroadaptive Technologies. It also suggests ways how such technology could be implemented, inspiring studies that are not aiming at evaluating BCIs, but furthermore investigating how this approach can be integrated into technologies to support people in their daily life. Neuroadaptive systems capable of inducing user-specific adaptation to the system's behavior might not only improve reading experience but could be particularly useful in any workload-sensitive context, such as personalized learning environments or online tutoring systems, leading to new ways for technology to present information to us.

Study 3: Monitoring neural correlates of human moral assessments for future human-machine value alignment

5.1 Motivation and summary of the study

As introduced in 2.7.7 the scope of potential areas of applications (in HMI) for passive BCI technology could be broadened if even more mental states were assessable in a user on a single-trial basis through passive BCI technology.

Effective moral value alignment between machines and humans is widely regarded as essential for successful future human-machine interaction (HMI). For this aim, autonomous moral agents (AMAs) must be enabled to learn about how morally right or wrong their behavior is evaluated. These moral evaluations could come from the human user who is immediately affected by the agent's behavior. Potentially, neurophysiological measures reflective of human moral evaluation might help to achieve this goal.

In this preliminary study we examined the foundations of enabling an intelligent system to automatically assess information about its user's subjective moral evaluation through a passive BCI, while the user is interacting with the machine. Passive BCIs extract information with no need of awareness from the human user, cognitive overload or any distraction from the ongoing context is avoided. The approach is thus practically applicable in uncontrolled scenarios. Over time, the AMA would build up a moral experience by continuously updating a moral value representation through learning directly from the user's cortical responses to moral aspects of the context. In that way, an AMA could be morally aligned neuroadaptively towards a specific user.

To that end, the aim of this study was to investigate neural markers of human morality and whether a personalized passive BCI can be trained successfully, to distinguish between morally objectionable and neutral information, hence to be sensitive to moral evaluation as a user state. To date, some neurophysiological studies indicate that the ERPs of stimuli which are subjectively perceived as morally acceptable or unacceptable show significant differences in ERP components. However, it has not been investigated yet, whether these differences in the EEG signal can be used on a single-trial basis to train a passive BCI classifier model that is sensitive to human moral evaluation.

This investigation aims to shed light on the question if the output of a passive BCI can serve as such an indicator of moral evaluations and can therefore be employed as a reward function to reinforce the moral value representation in an AMA, potentially in real-time, to that of the perceiving user and provide further insight into the neurophysiology of human morality. This approach towards neuroadaptive human-machine moral value alignment could pose an important step towards improved and individualized HMI, as well as a basis for successful future human-machine collaboration.

In the presented study, a picture-based calibration paradigm was developed, containing pictures which were ranked as morally unacceptable, as well as morally neutral. We trained a passive BCI classifier model to distinguish between those two trial classes on a single-trial basis on the basis of recorded EEG data. Results of this approach showed low classification accuracies which were not better than random. We then transferred the same passive BCI classifier calibration approach to EEG data from two other studies (Leuthold et al. (2014) and Berkum et al. (2009)) which already recorded and investigated differences in EEG between moral properties on a neurophysiological ERP analysis level, where they found significant differences between morally acceptable and unacceptable trials on an averaged level. For both of these studies classification accuracies again turned out to be very low. The results of our study indicate that, to date and with state of the art classification approaches, it is not possible to train an individualized passive BCI classifier model to reliably distinguish between a user's subjective moral evaluations on a single-trial basis.

5.2 Theoretical background

5.2.1 Autonomous Systems

Intelligent computer technology has become indispensable in modern society. Through advances in *artificial intelligence* (AI, Russell and Norvig, 2010) – the ability of a computer or computer-controlled robot to perform tasks commonly associated with

intelligent beings – machines handle mounting complex tasks with more precision and in less time than humans. Such *intelligent systems* have the "capacity to acquire and apply knowledge in an "intelligent" manner and have the capabilities of perception, reasoning, learning, and making inferences (or decisions) from incomplete information" (p.4, Karray et al., 2004). The capabilities of intelligent systems to learn how to solve problems successfully from more complex data have enhanced through advances in *machine learning* (ML, Bishop and Nasrabadi, 2006), *deep learning* (Goodfellow and Bengio, 2016) in particular.

To this day the tasks intelligent systems are capable to perform are focused on highly specific domains (i.e. playing chess (Newborn, 2012) or image classification (Yang et al., 2002)). However, the scope of tasks intelligent systems are employed to is mounting, progressively resulting in enhanced interaction between humans and technological systems through *ambient intelligence* (Tapia et al., 2010), e.g. in human-agent teamwork in eHealth (Ball et al., 2010) or personalized learning assistants (Canbek and Mutlu, 2016). In addition, intelligent systems are increasingly entrusted to autonomously respond to their environment, independent of direct instruction, as e.g. autonomous cars, assisted living technology or vacuum cleaners (Kleinberger et al., 2009; Baber et al., 2005; Saffiotti and Broxvall, 2005). Such *autonomous systems* can detect changes in their environment in real-time, meaningfully segment this information to learn about underlying patterns in order to effectively make a decision and act accordingly.

Regarding successful interaction and collaboration between humans and autonomous systems the question arises, in accordance to which moral values an autonomous agent obtains its decisions. Thus, as machines act with reduced human supervision, design principles for intelligent systems are needed that are non-conflicting with human moral goals (Russell, 2016). Thence, different disciplines of research have been concerned with the question of how a representation of human moral values can be implemented in autonomous machines and how such a machine can integrate those moral values into its decision-making process. The *value alignment problem* (Soares and Fallenstein, 2014) – a specification of a moral value representation for machines reflecting human values – can be structured in three sub-problems (Yampolskiy, 2019). The first of the sub-problems is concerned with personalized moral value extraction from individual persons, while the other two deal with an integration or combination of moral values of humanity as a whole. Yampolskiy (2019) argues that research should focus on implementing a "solution which optimally aligns with individual preferences of each user" (p.1), as such a system would "be easier to design, implement and safeguard" (p.2) due to overall lower complexity than trying to implement the "likely impossible to solve problem

of multi-agent value aggregation” (p.4).

5.2.2 Artificial Moral Agents

Artificial moral agents (AMA, Allen et al., 2000) are machines that possess the ability to reason morally. Thus, *autonomous* AMAs are systems equipped with a moral value representation which influences the process of selecting one alternative of a set of possible actions. Different areas of research have developed theories and rudimentary implementations of AMAs through different approaches (see e.g. Allen et al., 2005; Wallach and Allen, 2009).

In *top-down* approaches, machines are provided with explicit decision rules and parameters for selection of morally appropriate actions. Such AMA architectures are often based on explicit formulations of *normative systems* (e.g. utilitarianism, Kantian deontology, Ten Commandments, Asimov’s laws of robotics) and are mostly implemented through logic programming. One problem for the implementation of AMAs through top-down approaches is that there exists no consensus among different groups and cultures of humans – and even within individuals themselves – regarding which existing normative system would be overall preferred to be implemented in an AMA. In case a set of different moral normative systems gets implemented in the AMA at once, it can become impossible to reach a moral decision due to contradicting moral policies. Additionally, according to the *frame problem* (Dennett, 1984; Shanahan, 2000) it is often difficult to know which information is (ir)relevant in a situation and therefore should(n’t) be considered for moral decision making (Wallach, 2010).

AMAs can be implemented through a *bottom-up* approach, as well. Here, data is provided to the machine, representing situations and the morally preferred solution for each case. To derive a model of the underlying human morality, ML algorithms are applied to find patterns – in form of transitive, systematic relations between events and moral evaluations – in the data. Hence, the representation of human morality in this approach is a non-static set of rules emergent by learning, rather than a set of predefined rules and parameters. An example for the bottom-up approach is the *Moral Machine* (Awad et al., 2018) project where crowd-sourced data was used to train an algorithm to decide about moral dilemmas in the context of self-driving cars. A problem posed by bottom-up approaches is the absence of transparency, as often it can be challenging to reconstruct a system’s moral reasoning (Misselhorn, 2020). Further, as required by current ML approaches, the collection of large amounts of training data implies high effort and time. Many people must be questioned regarding their views on specific morally laden situations for a variety of topics, to achieve a close-to-complete representation of human moral

values.

As there are relevant limitations to the two approaches above, and currently no other concepts of great significance are available, it is largely proposed (Allen et al., 2005; Misselhorn, 2018) that a *hybrid approach* would prove most effective. Here, a predetermined fundamental framework of moral decision-rules (top-down) that is subsequently customized to particular moral scenarios through learning processes (bottom-up) combines the advantages of both methods. A hybrid approach could maintain transparency about the underlying processes the system bases its moral decisions on.

5.2.3 Machine Morality for Human-Machine Interaction

A successful future collaboration between humans and machines is based on a form of *machine morality* (Wallach et al., 2008) - machines inhabiting a capacity to assess and respond to moral challenges autonomously – as discussed above. *Value alignment* between robots and humans is widely regarded as advantageous for successful HMI.

Goetz and Kiesler (2002) e.g. found that personality traits that are displayed by a robot during HMI can influence the user's willingness to comply with the robot, such as AI assistants like Apple's Siri, Google's Assistant or Microsoft's Cortana. Additionally, trust is an important factor that mediates interactions between intelligent systems and users (Ghazizadeh et al., 2012). Further, social relatedness seems to be a factor influencing task performance (Tapus and Matarić, 2008).

Above that, in deliberations concerning AI safety (Amodei et al., 2016)– particularly in the prospect of sophisticated AI systems potentially leading to *superintelligence* (Shanahan, 2019) – it is widely proposed to aim for intelligent machines equipped with a moral value representation benevolent towards humans (V. C. Müller and Bostrom, 2016; Sarma and Hay, 2017). In point 10 of 23 of the Asilomar Principles (2018) - a set of ethical guidelines for AI developers under the aspect of AI safety declared by AI experts at the Future of Life Institute's Beneficial AI Conference 2017 - it is stated that „highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation”.

In humans, "morality is primarily learned, built up through experience, trial and error" (p.152, Allen et al. (2005)). To provide a similar process to a machine, Allen et al. (2000) propose developing AMAs through a form of simulated childhood. It would consist "of a training period involving feedback about the moral acceptability of actions" (p.258) through an associative learning scheme, e.g. artificial neural networks. The AMA would work towards an acceptable level of moral behavior by

going through a series of learning situations created by humans. Quoting Alan Turing's idea of the *child machine* (Turing and Haugeland, 1950), Allen et al. (2005) argue that putting the computer through an *education* process, as a child undergoes, the machine can acquire sensibility towards moral actions and evaluations similar to humans. Through an ongoing *experience* process of trial and error, the machine can then update and refine the learned model of human morality through feedback allocation during interaction with humans. The mentioned *education* process can be implemented through hybrid approaches, but there is currently no mechanism to implement the *experience process*, which would require an "endless flow of fresh real world data that expand the domain the system peruses in its quest" (p.152, Allen et al., 2005).

Approaches for implementation of actual AMA architectures, as introduced in ii, are highly interdisciplinary with relevant limitations for now, but as pointed out, a combination of bottom-up at top-down architectures seems promising. Similar to Yampolskiy (2019), Misselhorn (2018) proposes, that for the bottom-up part, a moral system might be designed "to adapt to the *individual* moral value profile of [a specific] user" (p.166). This would entail representing interindividual differences in how people weigh moral values, as well as intraindividual changes in importance of certain moral values in different phases of life (Misselhorn et al., 2013). To implement such a system, Misselhorn (2018) suggests the machine presents scenarios of morally relevant situations to the user to then develop a moral value representation based on the user's reaction to those scenarios. The individual moral model can then be continuously trained and adapted through interaction between the individual user and the machine where moral values are weighted and recognized. Hereby, the system could gain a capacity to incorporate a form of human-imitating morality into its actions, learning over time to decide what is morally good and bad. Especially in contexts where a single user is affected by the system's decisions, "the system's decisions should come as close as possible to the decisions that the user would make in corresponding situations" (p.8, Misselhorn (2018)). Here, the system would function as a "*moral avatar*" (p.8) of its user to improve human-machine collaboration.

5.2.4 Approaches for Human-Machine Value Alignment

Different approaches have been proposed to accomplish value alignment. E.g. the *Quixote program* (Riedl and Harrison, 2016) is aiming to teach morality to robots. By reading narratives, robots are trained about acceptable sequences of events and successful ways to behave humanly moral. Here, in accordance to the text's contents, the machine starts acting morally, as a plot graph is translated into a tra-

jectory tree from which a reward function is derived for Q-learning (Watkins and Dayan, 1992). Riedl and Harrison (2016) argue that this approach is suitable for agents with limited application domains, but a need to interact with humans to achieve their purpose. This story reading approach can serve towards value alignment by training a moral agent in its moral education phase. This approach has its limitations though, if the goal is to align an agent's moral value representation to that of an individual human and adapt this representation over time, as changes in the user's moral preferences can occur. Furthermore, it is questionable whether the approach really can represent a *complete* moral model of a specific person, as it appears complicated to compile it in narratives.

Currently, there are no feasible means for everyday practice to continuously collect data on the human user's moral evaluation of an agent's moral behaviour. Dobrovsky et al. (2017) asked participants to observe and evaluate actions of intelligent systems in a serious gaming context. The evaluation, communicated through button presses, was used as a context-sensitive score for a reward function to enable reinforcement learning (Kotsiantis, 2007; Lample and Chaplot, 2017). This approach of active evaluation, however, is cumbersome and demanding for the observer, poses a potential source of error in highly complex environments and takes a lot of time to generate larger data sets. Further, active moral evaluation through the user might interfere with the very process we are aiming to measure, as the user actively allocates cognitive resources towards the communication of the moral evaluation and away from their primary task or activity. Subjective moral evaluation in real-time would be necessary, though, if the AMA's moral value representation should be adaptively aligned to that of its user. Here, an unobtrusive measurement, indicating the user's moral assessment towards the machine's actions, without distraction from the task at hand, would be helpful. Misselhorn (2020) proposes to "classify the evaluations of the scenarios by the user with the help of an emotion-recognition software" (p.18). Modern approaches to directly assess information from brain processes can provide an additional, and probably advanced, solution for the problem of automatic user-state detection.

5.2.5 Neurophysiology of Human Morality

Moral reasoning and the acquisition of morality in humans has been investigated in several neurophysiological EEG based studies which found several ERP components reflective of cognitive moral evaluation. ERPs' waveforms are described according to latency (in ms) after stimulus onset and amplitude (N - negative, P - positive), e.g. P300 as described in section 2.4.1. Lahat et al. (2013) showed that compared to typical (semantic) deviations, moral violations produced greater N200

amplitudes. Loke et al. (2011) found associations between the P300 and a *late positive potential* (LPP - typically starting 300-500 ms after stimulus onset) as responses to moral scenarios and prosocial behaviors. Leuthold et al. (2014) saw a fronto-central LPP when participants were presented morally unacceptable statements or behaviors. The LPP was found by Berkum et al. (2009) as well, alongside an N400, in response to morally objectionable statements. Yoder and Decety (2014) also found that differences in moral evaluation were reflected in differential amplitudes in the LPP, as well as in the N100 and N200, with the right inferior parietal cortex and medial prefrontal cortex as potential locations of cortical areas, sensitive to those moral evaluations. Event-related analyses have the advantage of providing information to specific events in the context directly and typically provide a good signal-to-noise ratio. Frequency based analyses (e.g. Fumagalli et al., 2011) would provide continuous information about cognitive states of morality, but were not investigated in this study.

As detailed above, some aspects of a human's internal state are reflected in measurable, neurophysiological changes and can provide insight into the internal psychological state, i.e. into the user's cognition, emotions, and intentions. The presented results from moral research demonstrate that aspects of moral cognition can be measured by EEG recordings. Later components of the ERP, like the LPP, appear to be good candidates for a general correlate of moral value violations, while its specificity and its cortical sources are little investigated.

5.2.6 Utilizing Passive BCIs for Neuroadaptive Human-Machine Value Alignment

As introduced in 2.6, passive BCIs extract information with no need of awareness from the human, cognitive overload or any distraction from the ongoing context is avoided. The approach is thus practically applicable in uncontrolled scenarios, e.g. in scenarios where humans interact with machines (personal computers, robots, avatars). Based on these experiences, cortical correlates of subjective moral evaluation arising during human-machine interaction could be assessed in a similar fashion and used for personalized human-machine moral value alignment, which is the main motivation for this preliminary study.

For effective human-machine moral value alignment, AMAs must be enabled to learn about how morally right or wrong their behavior is evaluated. These moral evaluations could come from the human user who is immediately affected by the agent's behavior. Potentially, it might be possible to employ neurophysiological measures reflective of human moral evaluation to achieve this goal. An intelligent system could then be enabled to automatically assess information about its user's subjective moral evaluation through a passive BCI, while the user is interacting with

the machine and its environment. In that way, an AMA could be aligned neuroadaptively. Over time, it would build up a moral experience by continuously updating a moral value representation through learning directly from the user's cortical responses to moral aspects of the context.

One essential foundation for realization of such neuroadaptive moral alignment would be that cortical responses reflective of moral judgment can be identified by a classification scheme in real-time on a single-trial level with sufficient reliability. Hence, in this study we aim to investigate whether a personalized passive BCI can be trained successfully to distinguish between morally objectionable and neutral information. The data collected in our experiment therefore will be processed by a passive BCI to determine whether classification accuracies are reasonable enough to serve as a reliable indicator of subjective moral evaluation. Adequate classification accuracy is of high importance if the moral value representation in an AMA should be effectively adapted depending on the classifier's output.

The specific research questions of this preliminary investigation are detailed in the following section.

5.3 Research questions

If a neurophysiological measure should serve as an indicator of human moral evaluation, brain responses towards morally acceptable, neutral or unacceptable information must exhibit measurably significant differences. Results in recent studies investigating neurophysiology of human morality through EEG measurement are not in unison. Nevertheless, they show a sensitivity to moral evaluation in different components of the ERP, with a focus on late potentials, like the LPP. As a first research question of this study, I will investigate which components of the ERP can be used reliably as an indicator for subjective moral evaluation:

- **Research question 1:** Does the perception of morally objectionable information evoke an ERP that shows a significantly different morphology compared to neurophysiological responses evoked by morally neutral information?

Further, measurable neurophysiological differences in moral evaluation should be classifiable by a passive BCI in real-time on single-trial level. This constitutes the second research question posed in this investigation:

- **Research question 2:** Can moral objection or neutrality be detected in single-trial by a passive BCI?

To answer this question, the data collected in the presented experiment was processed by a passive BCI, trained to discriminate between information an individual human finds morally objectionable, or neutral, correctly classifies the information on new data significantly better than random (Mueller-Putz et al., 2008). Adequate classification accuracy is of high importance if the moral value representation in an AMA should be effectively adapted depending on the classifier's output. The general aim of this study is to investigate, whether the user state of moral evaluation can be assessed through a passive BCI - not only for moral value representation in AMAs, but also in other applications and contexts. This would broaden the versatility of passive BCIs on a general level.

5.4 Methods

5.4.1 Participants

Four subjects took part in this preliminary study, with an average age of 33.5 years (SD = 7.16 years). Two of these subjects were female. All subjects' vision was normal or corrected to normal. Prior to the experiment, subjects gave their written informed consent to participate in the study. The experiment was conducted along a guideline ethically approved by the local ethics committee.

5.4.2 Apparatus

We recorded EEG from 64 active electrodes using an EEG system from Brain Products (ActiChamp, Brain Products GmbH, Gilching, Germany). Electrodes were positioned according to the extended international 10/20 system, with ground electrode at AF2 and reference at FCz. Electrode impedances were kept below 10 k Ω . EEG data was recorded using the Brain Vision RDA (Brain Products GmbH, Gilching, Germany) and the LabRecorder (Delorme et al., 2012). The native sampling rate was 500 Hz. The MATLAB embedded toolbox EEGLAB (Delorme and Makeig, 2004) was used for EEG data analysis. All BCI classification and model application were performed using the open-source toolbox BCILAB (Kothe and Makeig, 2013). The experimental paradigm was implemented in MATLAB, using the Psychophysics Toolbox extensions (Brainard and Vision, 1997; Kleiner et al., 2007).

5.4.3 Stimuli

It is the aim of this experiment to investigate a participant's neurophysiological response towards a stimulus in dependence of the independent variable moral acceptability. As stimuli we presented standardized pictures from the *Socio-Moral Image Database* (SMID, Crone et al., 2018). This image database contains 2.941 freely available pictures with ratings (scale from 1-5) for *moral acceptance*, *valence* and *arousal* for each picture from 2.716 participants. Based on these rating, pictures for three distinct experimental categories were selected:

1. **moral:** *morally bad* pictures
2. **affect:** *morally neutral* pictures with valence and arousal ratings that do not significantly differ from those in category 1. (t-tests results: valence: $t(199)=0.92$, $p=0.360$, arousal: $t(199)=-0.303$, $p=0.762$) and are therefore morally neutral, but evoke *negative affect*
3. **neutral:** overall *neutral pictures* regarding morality, valence and arousal

We chose this distinction to be able to dissociate responses evoked from moral violations from those evoked by negative affect only, as moral violations are often accompanied with negative emotions. The average ratings for moral acceptance, valence and arousal for pictures from each experimental category are presented in table 5.1. Overall participants were presented 100 pictures from each category - 300 pictures in total. For example pictures from the three picture categories please see Appendix 4 to 6.

Table 5.1 Means and SDs for the three different stimuli categories on the moral acceptance, valence and arousal ratings scales from the SMID

	arousal		moral		valence	
	mean	SD	mean	SD	mean	SD
(1) morally bad	3.62	0.45	1.88	0.34	1.56	0.31
(2) morally neutral . affect bad	3.57	0.46	2.72	0.21	1.87	0.19
(3) neutral regarding morality and affect	2.97	0.22	3.01	0.12	2.99	0.25

5.4.4 Procedure

Each participant was presented all 300 pictures in randomized order in blocks of 15 to 20 pictures. Following the presentation of a white cross-hair (duration 3 seconds), each picture was presented in the center of an otherwise black screen for a duration of 3 seconds with an inter stimulus interval (ISI) of 3 seconds (3 seconds cross-hair presentation).

To ensure that participants kept their attention on the content of the presented pictures, but without pointing towards moral evaluation, we asked participants to briefly answer a question after each block regarding the pictures just presented (e.g. “Which of the pictures you have just seen did you associated most with the feeling of anger/joy/disgust etc.? Why? What was displayed in the picture?”). This question after each block only varied by the emotion that was asked for, which was chosen randomly from a list of 20 negative and positive emotions. Participants gave their answer towards the question via keyboard input. After every fourth block participants were allowed a self-paced break. The overall duration of the experimental paradigm was approximately 50 minutes. An overview of the experimental procedure and timeline is given in figure 5.1 and the experimental setup in the laboratory is shown in 5.2.

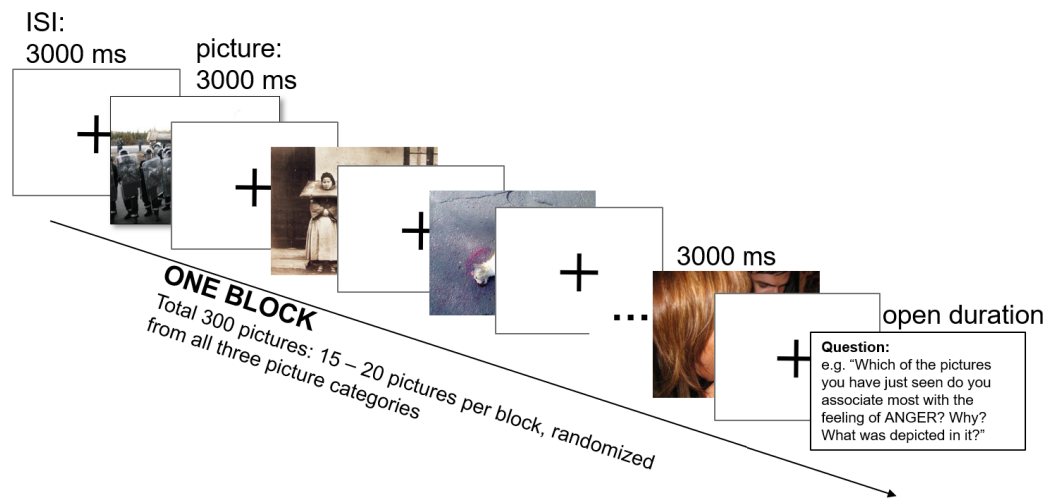


Fig. 5.1 Experimental timeline of the study presenting pictures from the SMID database. 300 pictures in randomized order in blocks of 15 to 20 pictures were presented to the participant. Each picture was displayed for 3 seconds with an ISI of 3 seconds during which a cross-hair was shown. After every block, participants were asked a question regarding the pictures of the block to which they give their answer via keyboard

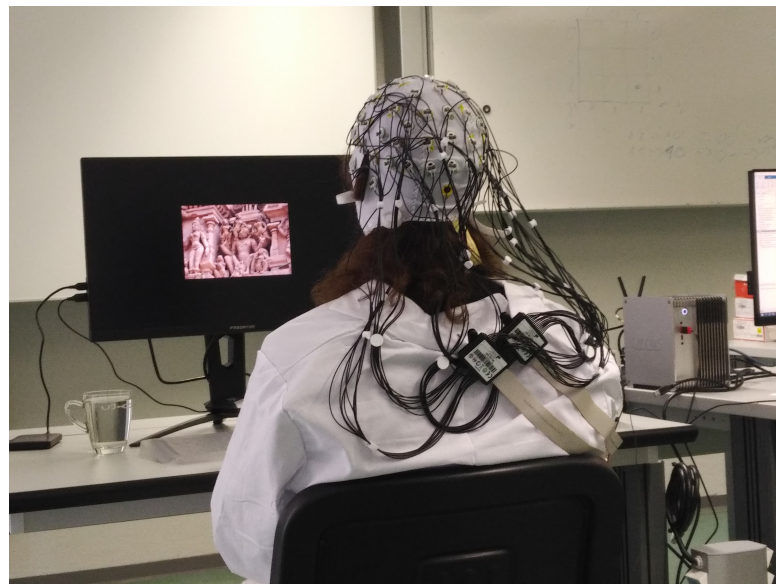


Fig. 5.2 Experimental setup for the SMID paradigm in the laboratory

5.5 Data Analysis

5.5.1 ERP analysis

EEG data recorded during presentation of the SMID pictures was subjected to an ERP analysis. The aim of this approach was to review whether the operationalization of the experiment produced ERPs which are conform with literature. ERPs of the investigated moral picture features were compared to ERPs from studies which had investigated the same moral features before. Raw data from the overall four participants was preprocessed by applying a bandpass filter from 0.5 to 40 Hz. Then data was re-referenced to the common average reference. Epochs of 1.2 seconds length were extracted starting 200 ms before picture presentation. A baseline correction was performed using the 200 ms before stimulus onset.

Grand average ERPs were generated from all trials of all participants for morally objectionable, neutral and affectively negative pictures. These grand averages were plotted for electrode sites Fz, Cz, and Pz, comparing grand averages of the three different condition separately and all together. The respective ERP difference curves (1) morally bad minus morally neutral, 2) affectively bad minus morally neutral and 3) morally bad minus affectively bad) between the three trial classes were also plotted, including p-values resulting from per sample permutation tests (see section 3.2.1.2), as we further analyzed the differences between the three trial classes, by performing respective permutation tests with 5000 permutations for every sample in the ERP as described in section 3.2.1.3.

5.5.2 Classification on stimuli features

In BCILAB features were extracted through the built-in windowed means approach (Blankertz et al., 2011). In this approach data was first down sampled to 100 Hz and a bandpass filter from 0.1 to 15 Hz was applied. Epochs of 800 ms length were extracted starting at stimulus onset. For 14 consecutive time windows of 50 ms length between 50 and 750 ms after each picture presentation, the average amplitude was calculated. Classification was performed through LDA (Duda et al., 2006) regularized by shrinkage (Blankertz et al., 2011). A (5x5)-fold cross-validation was performed to generate an estimate of the model's online reliability.

5.6 Results

In this section results from ERP analyses will be presented first, followed by reporting classification results from three different collections of data: The first collection of data was gathered from the experimental paradigm (described above) designed for this study based on the pictures from the SMID database. The other two collections of data on which classification was performed came from two studies of other research groups who investigated differences in EEG activity between morally acceptable and objectionable information and evaluation. These studies were aforementioned Berkum et al. (2009) and Leuthold et al. (2014).

5.6.1 ERP analyses

Grand average ERPs of the three different experimental stimuli classes (morally objectionable, morally neutral and affectively negative) were plotted and are shown in 5.3 (overview of all three stimuli classes). Please see Appendix section for pairwise plots with standard deviations (1 to 3). Comparing the ERP time courses of morally objectionable and morally neutral pictures (5.3), we see a slight early positivity (~ 70 ms) for morally objectionable trials after stimulus onset at Fz. We also find a negativity around 100ms and 270ms for morally bad trials at Cz. From around 350 ms onwards, we see a broad negativity in morally objectionable trials, compared to neutral trials on electrodes Fz and Cz. Between ERPs of morally objectionable and affectively bad trials (5.3), we find almost no differences in time courses, except for an early positivity (~ 70 ms) in morally objectionable trials at Fz and a small negativity for affectively bad trials ~ 270 ms at Fz. Finally, when comparing affectively bad and neutral trials, we find an early (~ 70 ms) negativity in affectively bad trials at Fz. We further find a negativity around 270ms for affectively bad trials as well in Cz. From around 350 ms onwards, we see a broad negativity in affectively bad trials, compared to neutral trials on electrodes Fz and Cz. The respective ERP difference curves between the two trial classes of the respective pairs of the three experimental stimuli classes are shown in figure 5.4. P-values of samples with significant permutation test outcomes are indicated as gray bars in the background. The results indicate that there are no significant differences between ERP curves for any combination of the stimuli classes at one of the electrodes Fz, Cz or Pz.

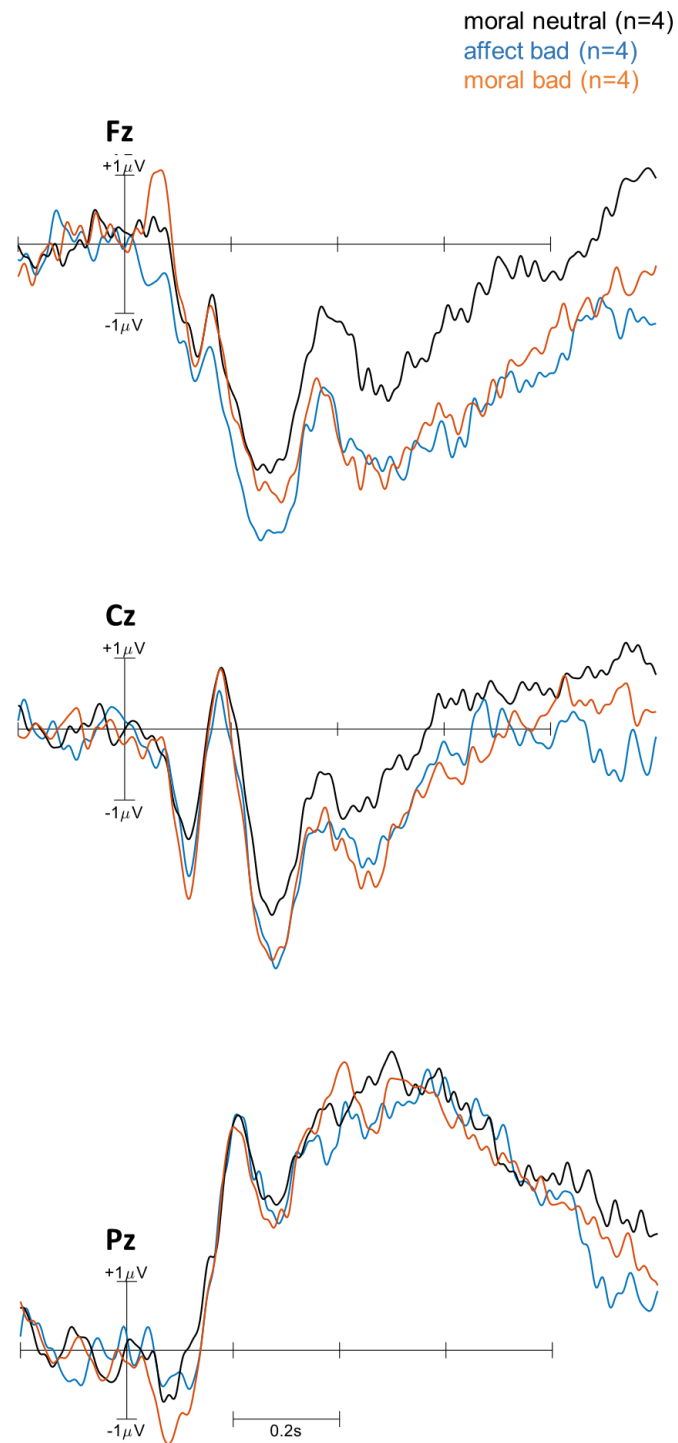


Fig. 5.3 Grand average ERPs (n=4) of morally objectionable, neutral and affectively negative trials from electrode sites Fz, Cz and Pz from the conducted experiment

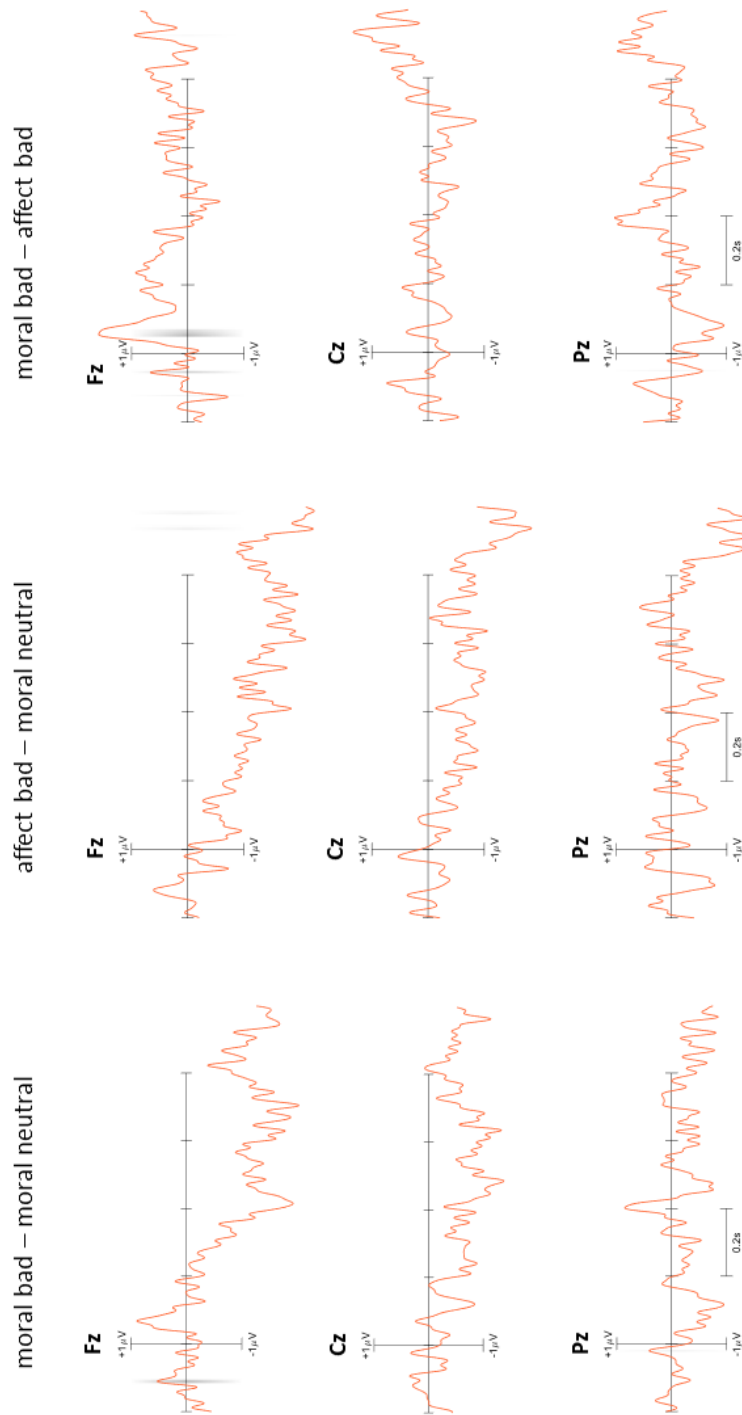


Fig. 5.4 Grand average difference curves of the ERPs (1) morally bad minus morally neutral, 2) affectively bad minus morally neutral and 3) morally bad minus affectively bad) at electrode sites Fz, Cz and Pz. Gray bars behind the plot indicate significant p-values from the permutation tests that were performed for each sample.

5.6.2 Classification analyses

In this section I firstly present classification results from the four subjects who completed the paradigm designed especially for this preliminary study – based on the above introduced pictures from the SMID database. These classification results showed (see below and table 5.2) that classification on these preliminary data was not better than random for any participant when we classified between morally objectionable and neutral trials. We concluded from these results that potentially the operationalization of our paradigm was not suited to evoke strong enough effects of moral evaluation to successfully make a distinction between moral objection and neutrality. Therefore, we asked authors of two other studies to share their data with us, so the classifier models could be trained on their data as well. In both of those studies, EEG data of morally acceptable and objectionable data was compared on grand average ERP level – including findings of significant differences between these two experimental conditions on a group level. Data from both these studies was kindly provided to us for classification analysis. Hence, classification could be performed in the same way as on our data on 24 datasets from an experiment performed by Leuthold et al. (2014) and on 43 datasets from a study conducted by Berkum et al. (2009). Findings on the grand average ERP level from those two studies as well as our classification results of these two data collections are reported at the end of this section.

5.6.2.1 Classification of data from the SMID pictures paradigm

Classification error rates obtained from the SMID pictures paradigm are shown in table 5.2. We tested whether the obtained classification performance of the individual classifier models was better than random classification (see section 2.5.3). The result of this check showed that the classification of none of the predictive models was better than random.

Table 5.2 Estimated classification results of the SMID picture based moral evaluation paradigm for all four subjects. Obtained error rates (ER) from (5,5) cross validation in percent and their standard deviations (SD), resulting from variance in the classification error estimated from each run of the cross validation are reported.

Subject	ER (SD)		
	moral vs. neutral	affect vs. neutral	moral vs. affect
1	48 (9.3)	51.5 (7.6)	44 (5.2)
2	46 (5.8)	39.5 (10.4)	50.5 (9.9)
3	58.5 (6.0)	48 (3.3)	57.5 (10.3)
4	48.5 (2.9)	48 (10.2)	48 (14.2)
Average	50.3 (6.0)	46.8 (7.9)	50 (9.9)

5.6.2.2 Classification on data from Berkum et al. (2009)

In their study van Berkum et al. (2009) created an experimental paradigm in which they presented statements to male participants who had one of two different value systems – half of participants were strict-Christians and the other half non-Christians. The last word of each presented statement would determine, whether the statement was consistent or inconsistent with the specific participant’s moral value system (e.g. “I think that euthanasia is acceptable OR unacceptable.”). Each participant read 90 of the statements with the critical evaluative word at the end – half of which were consistent and the other half inconsistent with the respective moral value system of the participant. While each of the overall 43 participants (21 strict-Christians and 22 non-Christians) read these statements, EEG was recorded from 32 electrode sites. The recorded EEG data was then subjected to ERP analysis for each of the experimental participant groups. For trials in which personal moral values were violated, Van Berkum et al. found a small N400 effect, as well as positivities around 200 to 250 ms. The LPP (500 to 650 ms) was found as well, in response to morally objectionable statements. These results were found for both groups of participants - strict-Christians and non-Christians.

For data from both experimental groups (non-Christian and strict-Christian) we applied the same classification scheme as above to train a classifier to distinguish between morally objectionable and acceptable trials. For participants from the strict-Christians group the average classification error rates was 47 percent (SD = 9.4). For non-Christians the average classification error for was 52.1 percent (SD = 10.9). All error rates are shown in table 5.3.

Table 5.3 Estimated classification results of data from Berkum et al. (2009) for all 43 subjects with morally acceptable and objectionable trials as classification classes. Obtained error rates (ER) from (5,5) cross validation in percent and their standard deviations (SD), resulting from variance in the classification error estimated from each run of the cross validation are reported.

Subject	ER (SD)	
	Strict-Christians	Non-Christians
1	47.8 (3.0)	53.3 (11.5)
2	44.4 (10.4)	56.7 (14.9)
3	47.8 (10.1)	50.0 (11.8)
4	53.3 (16.9)	54.4 (4.7)
5	46.7 (13.9)	51.1 (17.3)
6	55.6 (11.1)	50.0 (13.0)
7	36.7 (5.0)	40.0 (12.7)
8	43.3 (8.2)	65.6 (10.7)
9	43.3 (12.7)	54.4 (9.1)
10	45.6 (2.5)	51.1 (13.8)
11	37.8 (13.3)	51.1 (13.8)
12	50.0 (11.8)	55.6 (14.7)
13	55.6 (7.9)	47.8 (6.3)
14	36.7 (5.0)	48.9 (7.2)
15	46.7 (9.3)	44.4 (5.6)
16	43.3 (7.2)	63.3 (10.1)
17	57.8 (6.3)	53.3 (10.8)
18	40.0 (12.0)	43.3 (14.9)
19	52.2 (15.0)	56.7 (6.1)
20	51.1 (4.7)	48.9 (8.2)
21	52.2 (10.1)	45.6 (12.7)
22		60.0 (10.7)
Average	47.0 (9.4)	52.1 (10.9)

Table 5.4 Estimated classification results of data from Leuthold et al. (2014) for all 24 subjects with morally acceptable and objectionable trials as classification classes. Obtained error rates (ER) from (5,5) cross validation in percent and their standard deviations (SD), resulting from variance in the classification error estimated from each run of the cross validation are reported.

Subject	ER (SD)
1	57.5 (5.2)
2	47.5 (18.5)
3	48.9 (15.7)
4	62.5 (13.3)
5	63.8 (5.2)
6	55.0 (2.8)
7	62.5 (10.8)
8	42.5 (16.2)
9	48.8 (10.3)
10	50.0 (17.1)
11	47.5 (7.1)
12	53.8 (10.5)
13	42.5 (6.9)
14	48.8 (9.3)
15	53.8 (21.0)
16	47.5 (7.1)
17	42.5 (2.8)
18	41.2 (8.1)
19	42.5 (19.0)
20	57.3 (12.1)
21	57.5 (12.8)
22	36.3 (11.2)
23	46.3 (15.1)
24	51.3 (8.2)
Average	50.3 (11.1)

5.6.2.3 Classification on data from Leuthold et al. (2014)

In (2014) Leuthold et al. conducted a study in which they presented participants vignettes, which contained socio-normative scenarios, which at the very end either contained morally acceptable or unacceptable information in the text. Data from 34 participants was obtained, recording EEG activity from 72 electrode sites. Partic-

ipants first read each scenario's context text and upon button press, the final target sentence with the critical word at the end was displayed word by word on the experimental screen. Each participant was presented 40 acceptable and 40 unacceptable items. The results of an ERP analysis of the EEG data showed a fronto-central LPP when participants were presented morally unacceptable statements or behaviors. This broad late positivity started already around 320 ms after stimulus onset, peaking around 500 ms. For data from 24 of the originally 34 participants we applied the same classification scheme as before to train a classifier to distinguish between morally objectionable and acceptable trials. The obtained average classification error rate was 50.3 percent (SD = 11.1). All error rates are shown in table 5.4.

5.7 Discussion

In this section, the results of this study are detailed and discussed. Subsequently, two limitations in the experimental setup of this study are detailed, followed by an outlook on future research where an alternative experimental setup is proposed.

5.7.1 General discussion of the results

In this preliminary study we investigated neural markers of human morality and whether a passive BCI can be trained successfully, to distinguish between morally objectionable and neutral information. Such a BCI would be novel in the sense that it could add the user state of subjective moral evaluation and interpretation to the mental states that are neurophysiologically assessable through a passive BCI. While other user states, such as mental workload, error perception or attention (see section 2.6.2) can already be measured and monitored through passive BCI technology, the assessment of a user's subjective moral evaluation could enable an intelligent system to automatically obtain information about its user and thereby e.g. enable effective moral value alignment between machines and humans which is widely regarded as essential for successful future HMI. Potentially, neurophysiological measures reflective of human moral evaluation might help to achieve this goal, and could broaden the versatility of passive BCIs in general.

For this investigation, a picture-based calibration paradigm was developed, containing pictures which were ranked as morally unacceptable, as well as morally neutral. We trained a passive BCI classifier model to distinguish between those two trial classes on a single-trial basis on the basis of recorded EEG data. ERP analyses showed no significant differences between the ERP curves of the two categories. Results further showed low classification accuracies (for morally neutral and objectionable trials) which were not better than random.

We then transferred the same passive BCI classifier calibration approach to EEG data from two other studies (Leuthold et al. (2014) and Berkum et al. (2009)) which had already recorded and investigated differences in EEG data between moral properties on a neurophysiological ERP analysis level. Both studies found significant differences between morally acceptable and unacceptable trials on an averaged group level. Classification accuracies resulting from training a passive BCI classifier model on the data of the two studies again turned out to be very low. Altogether, the results of our study indicate that, to date and with state of the art classification approaches, it is not possible to train a passive BCI classifier model to reliably distinguish between a user's subjective moral evaluations on a single-trial basis.

5.7.2 Limitations and restrictions

One restriction which all three experimental calibration paradigms investigated throughout this study have in common, is that the ground truth of the moral acceptability of the stimulus material is assumed, evaluated and/or averaged prior to the experiments. However, the actual moral evaluations of the specific individual subject whose moral evaluations are to be examined and reflected in the data are not assessed. Hence, the ground truth of the applied stimuli regarding their moral acceptability in the eyes of an individual subject is not known and rather vague in all three experiments. This may be problematic for categorizing stimuli into moral acceptability categories that correspond to the actual moral views of the specific person - the labeled data given to the classifier for training is not based on the specific human being whose moral evaluations are to be tested and reflected in the data. Overall, this unreliably labeled training data may lead to classifiers that do not perform better than random - as seen in the trained classifier models from all three experiments we investigated in this study.

In the presented SMID picture based paradigm, the moral content of every specific picture was rated beforehand by a bigger sample of people and these ratings were averaged to achieve an approximation of the moral evaluation (as well as affective evaluation) of each specific picture in the corpus – these average ratings can however strongly differ from the individual human's moral views.

In the experimental paradigm for the study performed by Berkum et al. (2009) assumptions about religious beliefs are applied to assign the individual participants to a group of moral views they are deemed to hold - though the actual moral evaluation of the presented stimuli might differ between individuals and depend on other factors than religious views - such as personal history or cultural upbringing.

Finally, in the study by Leuthold et al. (2014) assumptions about social situations and contexts are applied to determine, which scenario is categorized as morally acceptable. These assumptions again may be a good estimate of the average opinion within the general population about the presented situations, they however most likely can differ strongly for specific individual participants – again making the moral labelling of specific stimuli difficult and imprecise. In all three studies for the presented stimuli it can only be assumed whether a certain stimulus should be labeled as morally acceptable, neutral or objectionable – but we do not know if this is the true labeling for the specific human we are testing.

To approximate true labelling of the training data regarding the actual, subjective moral evaluation of a stimulus, a different experimental setup should be implemented and tested, where the ground truth of the stimuli is assessed within the experiment for labelling the data for classifier training. These labels should come

from the specific participants themselves.

Apart from the above restriction regarding the labelling of stimuli, perhaps the stimuli presented in two of the three examined experiments are too complex. The pictures from the SMID corpus e.g. depict complex, multifaceted scenes and the socio-normative scenarios presented in the experiment by Leuthold et al. (2014) also describe different aspects of scenes. Here, the approach by Berkum et al. (2009) seems the clearest, as only single sentences or statements are shown to the participants.

Finally, we employed LDA regularized by shrinkage for classification in this study. However, other classification techniques, such as Riemannian classification or Deep Learning approaches, have recently been shown to achieve high classification accuracies, particularly for ERP classification (Lotte et al., 2018; Congedo et al., 2017). These promising classification algorithms should be looked at and applied as well in future studies for investigating the single-trial detectability of moral evaluations.

5.7.3 Outlook and conclusion

Based on the limitations above, we would like to propose an alternative experimental setup: This experiment again serves to assess participants' responses towards short objectionable and acceptable statements regarding multiple moral values.

In a first part of the experiment, participants will complete an online questionnaire at home. Its purpose will be to assess individual participants' opinions towards different societal moral issues – these opinions will be used as the individual subject's ground truth for assigning stimuli to a specific category of morality. To that end, participants will read single statements containing an opinion towards a morally laden issue from different contexts. These statements will be structurally altered versions of original statements from the "Wahl-O-Mat" from the past eight years. The Wahl-O-Mat is Germany's official voting guidance program where voters can test how much their views regarding different issues are congruent to that of specific political parties. Examples of topics include retirement, health care, animal rights, the export of weapons, refugees, sovereign debt, fake news, and legalization of marijuana. Participants will be asked to rate on a 7-point scale how strongly they agree (7) or disagree (0) to the statement or if they are neutral (4). Each statement will be stated in two different ways, once with a positive opinion towards the issue and once with a negative. Overall participants will read and rate 90 statements in the two versions - 180 in total. For the second part of this experiment, only those participants will be invited who have a high consistency in accepting one version of a specific statement and rejecting the other version, or rating both as neutral. The

results of this part of the experiment will be an approximation of individual moral attitudes of specific participants towards the given statements.

The second part of the proposed experiment will be an EEG study with participants in the laboratory. The general experimental procedure will be the same as in the first part. The independent variable in this part of the experiment, will be the moral acceptability of each stimulus (objectionable, neutral and acceptable), as individually determined through the individual participant's ratings given in the first part. Participants will read the same statements again with the two versions on a screen, plus twenty new statements. But participants now will firstly only be presented the beginning of the statement, which is the same in both versions. Upon button press, the ending of the statement will be presented, which will define which of the two versions is presently read. Afterwards participants will rate again how much they agree to the read statement. Their answers will be compared to those they gave in the first part of the experiment. Only data of participants with a high consistency in ratings they gave in parts one and two of the experiment – examined through interrater reliability – will be further examined. Recorded EEG data will then be used to train a passive BCI classifier to distinguish between statements individual participants agreed or disagreed to. The classifiers and ERPs will be interpreted neuroscientifically to ensure their validity and learn about the underlying neurophysiology.

The above described new experimental setup - with classifier calibration data labeled by the specific participant themselves for an optimized representation of the ground truth of their specific moral attitudes - could be expected to result in workable individual classifiers, whose output could be used to assess participants' responses towards objectionable and acceptable statements regarding multiple moral values.

We hope that through the conduction of a new experiment implementing the experimental paradigm proposed above, we can demonstrate a passive BCI capable of assessing human moral evaluation, which as described before, could enable moral value alignment between a user and an AMA. More generally, subjective moral evaluation would be added as a new neurophysiological correlate of a mental states which can be utilized and assessed by a passive BCI for several applications and contexts.

Acknowledgements

I would like to thank Prof. Dr. Hartmut Leuthold, as well as Prof. Dr. Jos van Berkum for sharing their datasets with me.

CHAPTER 6

Conclusions

In this chapter I will close this thesis with a discussion and conclusion of the three studies presented in chapters 3 to 5. The detailed results of the individual studies, as well as the studies' specific limitations, are described and discussed in the respective chapters. Finally, I will present the limitations of this work and provide approaches and perspectives for future research.

6.1 General conclusions

This thesis contributes to the fields of passive BCI and Neuroadaptive Technology by three studies that examine distinct challenges for passive BCI applications in real-world contexts. Passive BCI holds enormous potential to improve HMI, by increasing e.g. safety, performance and personalization. Passive BCIs may further decrease e.g. errors and frustration in different applications and tasks through unobtrusive mental user state assessment and monitoring. However, before passive BCI technology can unfold its full potential towards revolutionizing the interaction between humans and machines (and possibly between humans themselves), several at present existing hurdles and challenges need to be overcome.

In chapter 3, the challenge of *subject- and session dependence* was investigated for passive BCI models that classify error potentials.

Today, when the predictive model of a BCI is trained to classify between classes of data, training data must be collected beforehand from the specific person, who later wants to use the trained model in an online application with new incoming data. This step of *subject-specific* training data collection is often time consuming and unpractical for most HMI contexts. I therefore investigated the training of a *subject-independent* predictive model: this model is trained on a group of other users' data and the obtained classifier model is then applied to online-testing data of a completely new person. This investigation was performed in the context of training an automatic classifier for (automation) error detection and correction in a real-world context.

We first trained *subject-specific* error detection classifier models with high classification accuracies and validated that these models base their decisions - error or no error detected - on cortical activity, specifically an area around the ACC that neuroscientifically is linked to error processing. We further showed that subject-specific classifier models from one recording session can be transferred to new data recorded on a different day - concluding, that subject-specific classifier models for error-detection are *session-independent*. We also trained an error-detection classifier on the basis of data collected from multiple persons and successfully applied this trained group classifier model to new testing data of a new subject - showing that the classifier is *subject-independent*.

Overall the results of this study showed, that we can train a classifier model which omits all classifier calibration time for the user. We further demonstrated that the number of electrodes this classifier model is trained on, can at least be halved from the initial 50 electrodes used in our study down to 24 electrodes - leading to reduced preparation time for hardware (sensors) setup. These two improvements to system preparation time bring automatic error-correction through passive BCI support one step closer to be applied in the real-world. In future research, the developed classifier should be tested for its performance in other tasks in which interaction or execution errors can occur during HMI. Above that, it should be investigated, if the classifier can detect other types of errors, such as e.g. observation errors. Moreover, automatic error-correction through passive BCI monitoring should be evaluated in different applications from a user experience perspective in order to identify needs and expectations from the user's perspective.

In chapter 4 I addressed the problem of *task-dependence*. In order to gather training data for calibrating a BCI, usually a so called calibration paradigm is developed: Here the subject completes a number of trials containing tasks that should evoke the cognitive signal of interest, as well as other trials where the signal is not evoked. On the basis of this calibration data, the classifier model is then trained to distinguish between data in which the neural activity of the cognitive signal of interest is present and when it is absent. Such a calibration paradigm is usually generated for every specific context, task and application where a BCI is trained for online application. Some tasks potentially are based on similar cognitive processes though, making it unnecessarily effortful to design a specific calibration paradigm for each use case or task.

Based on promising findings, I tested the application of a potentially *task-independent* mental workload calibration paradigm in a new task - a speed reading context, where subjects read texts of different difficulty and at different reading speeds. Firstly, subject-specific classifier models were trained after subjects completed this potentially task-independent mental workload calibration paradigm (which is based on

an arithmetic task). The resulting predictive models were then applied to monitor mental workload while subjects read text material of different difficulty and reading-speed.

We found that the resulting predictive values for mental workload were higher for difficult texts than for easy texts. Predictive values from texts presented at an increased speed were also higher than for those presented at a normal self-adjusted speed.

These results suggest that the mental workload based predictive model can be applied to the new tasks of reading texts at different difficulties and reading speeds, supporting the claim that it can indeed serve as a task-independent classifier for mental workload.

Finally, in chapter 5, I examined the challenge of finding *new neural correlates of mental states which can be utilized and assessed by a passive BCI* to generally broaden the versatility of passive BCI technology. The study aimed at exploring whether neural correlates of human *moral assessment* - a person's individual judgment of the moral quality of an event or scene that is presented to them - could be measured on a single-trial basis through a passive BCI using EEG data acquisition. A passive BCI capable of classifying and monitoring human moral assessments could e.g. be employed in moral user modelling for future human-machine (moral) value alignment. Moreover, it would add moral assessment to the possible user states which can be measured and monitored by a passive BCI. This would in general increase the versatility of application areas of passive BCIs.

For this investigation a calibration paradigm was developed, containing pictures which were ranked as morally unacceptable, as well as morally neutral. We trained a passive BCI classifier model to distinguish between those two trial classes on a single-trial basis based on recorded EEG data. Results of this approach showed low classification accuracies which were not better than random. We then transferred the same passive BCI classifier calibration approach to EEG data from two other studies which already investigated neurophysiological differences in EEG between moral properties. For both of these studies classification accuracies again were very low and not better than random.

The obtained results indicate that, to date and with state-of-the-art classification approaches, it is not possible to train a passive BCI classifier model to reliably distinguish between a user's subjective moral evaluations on a single-trial basis. We further noticed that all three experimental paradigms taken into account in this study assumed or estimated the moral ground truth of the provided stimuli - without ensuring or assessing that the moral assessments and stimuli categorizations correspond to the subjective moral beliefs and evaluations of the individual subjects who completed the studies. Hence, we propose an alternative experimental

setup, where for each subject, classifier calibration data is labeled by the specific subject themselves for an optimized representation of the ground truth of their specific moral attitudes.

Overall, the major results presented in this thesis serve to solve some of the previously identified challenges of passive BCI technology for HMI and mark a significant step towards the application of passive BCIs in real-world situations and contexts.

However, while these findings represent significant advances in terms of reducing or eliminating classifier calibration time, hardware preparation (via electrode reduction), and providing generalizability of classifiers that function independently of the specific task at hand, a bottleneck for real-world applications of passive BCI technology remains in the currently used EEG headsets and hardware - here, more advances in reduced electrode setups, as well as dry electrode systems that reduce hardware preparation time to a minimum, are required.

Furthermore, the introduction of passive BCIs and NAT into everyday life would bring ethical challenges (as already detailed in section 2.7.4) in its wake, which should be given significant importance in the development and deployment of these technologies.

Apart from these more general future challenges, other more particular limitations of the presented work and recommendations for future research are discussed in the following section.

6.2 Future research and perspectives

Even though the results from the first two experiments presented in this thesis are a promising step towards overcoming the current challenges of task-, session- and subject-independence in passive BCI technology, this thesis still leaves a handful of valuable research areas for future work, which will be pointed out as follows.

- **Real-world application in realistic scenarios:**

Both, the subject- and session-independent error detection classifier developed in chapter 3 and the task-independent classifier for mental-workload assessment investigated in chapter 4, should be investigated for reproducibility in more natural online application contexts with users who solve real-world tasks. Even though the error-detection classifier was already applied to provide real-time feedback to support the user by automatic error-correction in the experimental setup in 3.2.7, the classifier should be applied in more realistic everyday settings to gain further insights into its performance, robustness and potential issues that may arise from real-world application - such as

confounding mental states in complex task environments or social settings, as well as real-life noise in EEG signals.

Further, the users' needs should be considered as well to ensure efficient BCI communication based on user-centered design (Kübler et al., 2014). Thus, testing real-time error detection and correction through the passive BCI classifier in a multitude of tasks and scenarios could give conclusions e.g. to the kind of error correction users prefer and find helpful for improved HMI from a user experience perspective.

- **Investigation of universal classification:**

The presented studies demonstrated a session- and subject-independent classifier for error detection, and showed the task-independence of the mental workload classifier for the task of text reading (text difficulty and reading speed). A classifier that would show those features together - session-, task- as well as subject-independence - could be used for *universal* classification of a specific mental state (as e.g. error perception or mental workload). A universal classifier would be desirable, as it could be applied without any calibration data recording for the specific user and the classifier could be used in a variety of tasks and scenarios, where the mental state of interest occurs or changes within a user. To test for universality, e.g. of the error detection group classifier, we would still need to demonstrate that it can be applied in other tasks, as well as for other types of errors. To that end, the calibration data we collected in the study presented in chapter 3 could be combined with EEG data from other studies in which errors were induced and studied. These studies can differ in the performed task by the user and machine, as well as in the type of error that occurred or should be detected. Through calibration and classification on subsets of this obtained pool of EEG data the potential for an additionally task- and/or error-type independent classifier model for error detection could be investigated.

On the other hand for the task-independent mental workload classifier, subject-independence could be additionally investigated. For the mental state of mental workload however, subject independence could be difficult to show in addition to task-independence, due to interindividual differences of task-ability. While e.g. an arithmetic task (which currently is the calibration task applied in the calibration paradigm of the task-independent classifier - see section 4.3.3) is relatively difficult for one individual user, the task of text reading or mental rotation could be relatively easy for this particular user. Hence, for every task there are different difficulty levels within the task that induce different levels of mental workload in the user, but the overall span of mental

workload covered by the specific task within the full, general span of mental workload within a person differs between users, as demonstrated by Zhang et al. (2018). Due to these interindividual differences in task-ability, it could be difficult to train a task-independent group classifier model and transfer it to different users.

- **Parallel investigation of current challenges for passive BCI technology:**

Currently, many studies researching current challenges in passive BCI technology aim to implement a solution to single, isolated challenges one at a time. However, designing experiments that investigate different challenges in parallel may result in more efficient and more informative results for real-world applicability of passive BCIs.

In the first study of this thesis we e.g. not only investigated session- and subject-independence of the error detection classifier, but also explored the reduction of the employed number of EEG channels for data acquisition. The latter is an important aspect when considering usability factors for end-users who request minimal hardware setup preparation time for using BCI technology for support. As we included electrode reduction into our experimental setup, we could not only demonstrate session- and subject-independence of the classifier, but also, that it can still reliably operate with only half the original electrodes - this combination of findings (a plug-and-play classifier model plus a reduced number of required sensors) allows us to draw more realistic and optimistic conclusions about the real-world applicability of the investigated error-detection classifier.

Similarly, we could have applied the error-detection classifier to another HMI application in which errors occur, to investigate the classifier's applicability to other tasks and contexts - its task-independence.

- **Provide context information for mapping of events and detected mental state changes:**

Ongoing EEG recordings in real-world measurements inherently lack event markers that could link specific contextual events to the output of a passive BCI classifier. Hence, even if subject- or task-independent classification can be performed, in real-world situations time-locked events are rarely available to e.g. classify specific events at specific time points for the detection of errors. Thus it can be difficult to connect the outputs of a passive BCI classifier model with the situation or event which elicited it in real-time applications in the real world, to later make appropriate adaptations to the application a user is interacting with. Therefore, measures that can provide context and origin of events which evoke passive BCI detectable changes in user state should be

investigated. For instance a multimodal interface that allows for combined measurement of EEG data and eye tracking, could be employed as an indicator to where the focus of visual attention was, when e.g. an error potential was identified in the ongoing stream of analysed EEG data.

- **Employ classification algorithms that require larger amounts of calibration data and time:**

In (passive) BCI research to date the machine learning procedure of choice remains LDA - often regularized by shrinkage - (see section 2.5.3) due to its usefulness in small training samples settings. However, in recent years other machine learning methods and algorithms for classification of EEG signals in BCI have been developed and tested - for an extensive overview and review of these classification algorithms for EEG based BCI please see Lotte et al. (2018).

Deep learning methods for data classification are generally promising approaches for BCI classification, but their performance suffers greatly when only small sample sizes are available as training data. This however, usually is the case in BCI applications, where in a specific, time-limited session training data is collected from the individual BCI user. Lotte et al. (2018) particularly conclude that currently the greatest limitation for e.g. *Deep Neural Networks* (DNNs) "for EEG-based BCI is that such networks have a very large number of parameters, which thus requires a very large number of training examples to calibrate them" (p.35). They propose that one solution to this problem would be "to obtain BCI applications with very large training data bases, e.g., for multi-subject classification" (p.35). This solution can be provided by the approach for training of a session- and subject independent classifier model which we presented in chapter 3, which provides a much larger amount of training samples, as the group classifier for error-detection can be trained on data from numerous subjects that is collected over time.

The presented approach also provides a solution for another issue which Lotte et al. (2018) raise for DNN classification: due to their computational complexity, training and testing times are generally long - hours or days - which is not feasible for online BCI applications, where even a few minutes are too long for users to wait until the classifier is trained and ready for application. Lotte et al. (2018) propose that "designing DNNs that do not require any subject-specific training, i.e., a universal DNN" (p.35) would be an alternative. Here again, the subject-independent classifier we demonstrated in chapter 3 could provide an according solution, as the subject-independent error-detection group classifier can be trained in a single classifier calibration on the basis

of data from many subjects - at a time before a new user wants to apply the classifier model to assess their own brain activity.

For future research we therefore propose to train DNN classifiers on the calibration data from all subjects of the first study of this thesis (and possibly collect more data from more subjects beforehand).

Based on the research approaches proposed above, another more general recommendation for future (passive) BCI research emerges: the BCI field might seemingly profit from a **centralized repository for experimental recordings**, where researchers could be given access to data from other scientific groups that recorded neurophysiological (calibration) data that can be categorized by the investigated mental state (mental workload, error perception, emotions etc.), recording scenario (real-world, laboratory), electrodes (number, type and setup), tasks performed by the user (arithmetic, game play, text reading etc.), sampling rate and other feasible categories for data selection. Recording neurophysiological data is very time consuming and expensive. Therefore, such a repository would most probably prove to be very helpful and time efficient, as researchers could potentially find already existing data in the repository that would be suitable for investigating their research questions, as e.g. as detailed above, the subject-independent group classifier could be tested on other data from a different experiment where errors were induced as well, but a different task was performed by the subject.

Even though the initialization of such a repository may require a substantial amount of time and effort (and must be maintained and supervised), as well as some level of agreement between researchers about the data labelling structure and standardization, I strongly believe that such a hub for data collaboration and exchange could accelerate advances in passive BCI research and foster collaboration and knowledge exchange between international research groups and between scientific disciplines in the future.

References

- Acı, Ç. İ., Kaya, M., and Mishchenko, Y. (2019). Distinguishing mental attention states of humans via an eeg-based passive bci using machine learning methods. *Expert Systems with Applications*, 134, 153–166.
- Alchalabi, A. E., Shirmohammadi, S., Eddin, A. N., and Elsharnouby, M. (2018). Focus: Detecting adhd patients by an eeg-based serious game. *IEEE Transactions on Instrumentation and Measurement*, 67(7), 1512–1520.
- Alimardani, M., and Hiraki, K. (2020). Passive brain-computer interfaces for enhanced human-robot interaction. *Frontiers in Robotics and AI*, 125.
- Allen, C., Smit, I., and Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7.
- Allen, C., Varner, G., and Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12.
- Al-Nafjan, A., Hosny, M., Al-Ohali, Y., and Al-Wabil, A. (2017). Review and classification of emotion recognition based on eeg brain-computer interface system research: a systematic review. *Applied Sciences*, 7(12), 1239.
- Alsaggaf, E. A., and Baaisharah, S. S. (2014). Directions of autism diagnosis by electroencephalogram based brain computer interface: a review. *Life Science Journal*, 11(6), 298–304.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). *Concrete problems in ai safety*.
- Andreessen, L. M., Gerjets, P., Meurers, D., and Zander, T. O. (2021). Toward neuroadaptive support technologies for improving digital reading: a passive bci-based assessment of mental workload imposed by text difficulty and presentation speed during reading. *User Modeling and User-Adapted Interaction*, 31(1), 75–104.
- Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *2008 ieee international joint conference on neural networks (ieee world congress on computational intelligence)* (pp. 2390–2397).
- Anh, V. H., Van, M. N., Ha, B. B., and Quyet, T. H. (2012). A real-time model based support vector machine for emotion recognition through eeg. In *2012 international conference on control, automation and information sciences (iccais)* (pp. 191–196).

- Appriou, A., Ceha, J., Pramij, S., Dutartre, D., Law, E., Oudeyer, P.-Y., and Lotte, F. (2020). Towards measuring states of epistemic curiosity through electroencephalographic signals. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 4006–4011).
- Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., ... others (2016). Adaptive automation triggered by eeg-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment. *Frontiers in human neuroscience*, *10*, 539.
- Arico, P., Borghini, G., Di Flumeri, G., Colosimo, A., Pozzi, S., and Babiloni, F. (2016). A passive brain-computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks. *Progress in brain research*, *228*, 295–328.
- Arico, P., Borghini, G., Di Flumeri, G., Sciaraffa, N., Colosimo, A., and Babiloni, F. (2017). Passive bci in operational environments: insights, recent advances, and future trends. *IEEE Transactions on Biomedical Engineering*, *64*(7), 1431–1436.
- Asilomar, A. (2018). Principles.(2017). In *Principles developed in conjunction with the 2017 asilomar conference [benevolent ai 2017]*.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, 563.
- Baber, J., Kolodko, J., Noël, T., Parent, M., and Vlacic, L. (2005). Cooperative autonomous driving: Intelligent vehicles sharing city roads. *IEEE Robotics and Automation Magazine*, *12*.
- Baldwin, C. L., and Penaranda, B. (2012). Adaptive training using an artificial neural network and eeg metrics for within-and cross-task workload classification. *NeuroImage*, *59*(1), 48–56.
- Ball, M., Callaghan, V., Gardner, M., and Trossen, D. (2010). Achieving human-agent teamwork in ehealth based pervasive intelligent environments..
- Barry, R. J., Clarke, A. R., Johnstone, S. J., Magee, C. A., and Rushby, J. A. (2007). Eeg differences between eyes-closed and eyes-open resting conditions. *Clinical neurophysiology*, *118*(12), 2765–2773.
- Bell, T. (2001). Extensive reading: Speed and comprehension. *The reading matrix*, *1*(1).
- Benedetto, S., Carbone, A., Pedrotti, M., Le Fevre, K., Bey, L. A. Y., and Baccino, T. (2015). Rapid serial visual presentation in reading: The case of spritz. *Computers in Human Behavior*, *45*, 352–358.
- Berger, H. (1929). Über das elektroencephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten*, *87*(1), 527–570.

- Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., ... Craven, P. L. (2007). Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78(5), B231–B244.
- Berkum, J. J. V., Holleman, B., Nieuwland, M., Otten, M., and Murre, J. (2009). Right or wrong? the brain's fast response to morally objectionable statements. *Psychological Science*, 20.
- Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., ... Flor, H. (1999). A spelling device for the paralysed. *Nature*, 398(6725), 297–298.
- Bishop, C. M., and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer.
- Blanco, S., Garcia, H., Quiroga, R. Q., Romanelli, L., and Rosso, O. (1995). Stationarity of the eeg series. *IEEE Engineering in medicine and biology Magazine*, 14(4), 395–399.
- Blankertz, B., Acqualagna, L., Dähne, S., Haufe, S., Schultze-Kraft, M., Sturm, I., ... Müller, K.-R. (2016). The berlin brain-computer interface: progress beyond communication and control. *Frontiers in neuroscience*, 10, 530.
- Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K.-R. (2011). Single-trial analysis and classification of erp components—a tutorial. *NeuroImage*, 56(2), 814–825.
- Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., ... others (2010). The berlin brain–computer interface: non-medical uses of bci technology. *Frontiers in neuroscience*, 198.
- Bonaci, T., Calo, R., and Chizeck, H. J. (2014). App stores for the brain: Privacy and security in brain-computer interfaces. In *2014 ieee international symposium on ethics in science, technology and engineering* (pp. 1–7).
- Bonnet, L., Lotte, F., and Lécuyer, A. (2013). Two brains, one game: design and evaluation of a multiuser bci video game based on motor imagery. *IEEE Transactions on Computational Intelligence and AI in games*, 5(2), 185–198.
- Brainard, D. H., and Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, 10(4), 433–436.
- Brown, G. D., Yamada, S., and Sejnowski, T. J. (2001). Independent component analysis at the neural cocktail party. *Trends in neurosciences*, 24(1), 54–63.
- Buch, E., Weber, C., Cohen, L. G., Braun, C., Dimyan, M. A., Ard, T., ... others (2008). Think to move: a neuromagnetic brain-computer interface (bci) system for chronic stroke. *Stroke*, 39(3), 910–917.
- Byrne, E. A., and Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological psychology*, 42(3), 249–268.

- Canbek, N. G., and Mutlu, M. E. (2016). On the track of artificial intelligence: Learning with intelligent personal assistants. *International Journal of Human Sciences*, 13.
- Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., and Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280(5364), 747–749.
- Castermans, T., Duvinage, M., Cheron, G., and Dutoit, T. (2013). Towards effective non-invasive brain-computer interfaces dedicated to gait rehabilitation systems. *Brain sciences*, 4(1), 1–48.
- Chanel, G., Rebetz, C., Bétrancourt, M., and Pun, T. (2011). Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(6), 1052–1063.
- Chaouachi, M., Jraidi, I., and Frasson, C. (2011). Modeling mental workload using eeg features for intelligent systems. In *International conference on user modeling, adaptation, and personalization* (pp. 50–61).
- Chavarriaga, R. (2020). Standards for neurotechnologies and brain-machine interfacing [standards]. *IEEE Systems, Man, and Cybernetics Magazine*, 6(3), 50–51.
- Chavarriaga, R., and Millán, J. d. R. (2010). Learning from eeg error-related potentials in noninvasive brain-computer interfaces. *IEEE transactions on neural systems and rehabilitation engineering*, 18(4), 381–388.
- Chavarriaga, R., Sobolewski, A., and Millán, J. d. R. (2014). Errare machinale est: the use of error-related potentials in brain-machine interfaces. *Frontiers in neuroscience*, 208.
- Chen, J., and Song, X. (2016). Brain-computer interface in construction safety management: A quantitative framework. In *Construction research congress 2016* (pp. 2719–2729).
- Chen, X., Wang, Y., Nakanishi, M., Gao, X., Jung, T.-P., and Gao, S. (2015). High-speed spelling with a noninvasive brain-computer interface. *Proceedings of the national academy of sciences*, 112(44), E6058–E6067.
- Chuang, J., Nguyen, H., Wang, C., and Johnson, B. (2013). I think, therefore i am: Usability and security of authentication using brainwaves. In *International conference on financial cryptography and data security* (pp. 1–16).
- Collins-Thompson, K., and Callan, J. P. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004* (pp. 193–200).
- Congedo, M., Barachant, A., and Bhatia, R. (2017). Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3), 155–174.

- Constant, I., and Sabourdin, N. (2012). The eeg signal: a window on the cortical brain activity. *Pediatric Anesthesia*, 22(6), 539–552.
- Crone, D. L., Bode, S., Murawski, C., and Laham, S. M. (2018). The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes. *PLoS ONE*, 13.
- Crossley, S. A., Allen, D. B., and McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language*, 23(1), 84–101.
- Daly, I., Williams, D., Kirke, A., Weaver, J., Malik, A., Hwang, F., ... Nasuto, S. J. (2016). Affective brain–computer music interfacing. *Journal of Neural Engineering*, 13(4), 046022.
- Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, 19(4), 450–466.
- De Boer, R., and Dekker, S. (2017). Models of automation surprise: results of a field survey in aviation. *Safety*, 3(3), 20.
- de Bruijn, E. R., Schubotz, R. I., and Ullsperger, M. (2007). An event-related potential study on the observation of erroneous everyday actions. *Cognitive, Affective, and Behavioral Neuroscience*, 7(4), 278–285.
- Dehaene, S., Posner, M. I., and Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological science*, 5(5), 303–305.
- Dehais, F., Duprès, A., Blum, S., Drougard, N., Scannella, S., Roy, R. N., and Lotte, F. (2019). Monitoring pilot's mental workload using erps and spectral power with a six-dry-electrode eeg system in real flight conditions. *Sensors*, 19(6), 1324.
- Dehais, F., Dupres, A., Di Flumeri, G., Verdier, K., Borghini, G., Babiloni, F., and Roy, R. (2018). Monitoring pilot's cognitive fatigue with engagement features in simulated and actual flight conditions using an hybrid fnirs-eeg passive bci. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 544–549).
- Dehais, F., Peysakhovich, V., Scannella, S., Fongue, J., and Gateau, T. (2015). " automation surprise" in aviation: Real-time solutions. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 2525–2534).
- Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies* (pp. 73–83).
- Delorme, A., Kothe, C., Vankov, A., Bigdely-Shamlo, N., Oostenveld, R., Zander, T. O., and Makeig, S. (2010). Matlab-based tools for bci research. In *Brain-computer interfaces* (pp. 241–259). Springer.

- Delorme, A., and Makeig, S. (2004). Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1), 9–21.
- Delorme, A., Palmer, J., Onton, J., Oostenveld, R., and Makeig, S. (2012). Independent eeg sources are dipolar. *PloS one*, 7(2), e30135.
- Dennett, D. C. (1984). Cognitive wheels: The frame problem of ai. *Minds, machines and evolution*.
- De Visser, E. J., Beatty, P. J., Estepp, J. R., Kohn, S., Abubshait, A., Fedota, J. R., and McDonald, C. G. (2018). Learning from the slips of others: Neural correlates of trust in automated agents. *Frontiers in human neuroscience*, 12, 309.
- Dobrovsky, A., Borghoff, U. M., and Hofmann, M. (2017). An approach to interactive deep reinforcement learning for serious games..
- Donchin, E., and Cohen, L. (1969). Anticipation of relevant stimuli and evoked potentials: A reply to naatanen. *Perceptual and Motor Skills*, 29(1), 115–117.
- DuBay, W. H. (2004). The principles of readability. *Online Submission*.
- Duda, R. O., Hart, P. E., et al. (1973). *Pattern classification and scene analysis* (Vol. 3). Wiley New York.
- Duda, R. O., Hart, P. E., et al. (2006). *Pattern classification*. John Wiley and Sons.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. 605 third avenue. New York, USA: John Wiley and Sons.
- Durkee, K., Hiriyanna, A., Pappada, S., Feeney, J., and Galster, S. (2016). Multi-model approach to human functional state estimation. In *International conference on augmented cognition* (pp. 188–197).
- Ehrlich, S. K., and Cheng, G. (2018). Human-agent co-adaptation using error-related potentials. *Journal of neural engineering*, 15(6), 066014.
- Évain, A., Argelaguet, F., Strock, A., Roussel, N., Casiez, G., and Lécuyer, A. (2016). Influence of error rate on frustration of bci users. In *Proceedings of the international working conference on advanced visual interfaces* (pp. 248–251).
- Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with computers*, 21(1-2), 133–145.
- Fairclough, S. H., and Lotte, F. (2020). Grand challenges in neurotechnology and system neuroergonomics. *Frontiers in Neuroergonomics*, 2.
- Falkenstein, M., Hohnsbein, J., Hoormann, J., and Blanke, L. (1991). Effects of cross-modal divided attention on late erp components. ii. error processing in choice reaction tasks. *Electroencephalography and clinical neurophysiology*, 78(6), 447–455.

- Falkenstein, M., Hoormann, J., Christ, S., and Hohnsbein, J. (2000). Erp components on reaction errors and their functional significance: a tutorial. *Biological psychology*, 51(2-3), 87–107.
- Fan, J., Wade, J. W., Bian, D., Key, A. P., Warren, Z. E., Mion, L. C., and Sarkar, N. (2015). A step towards eeg-based brain computer interface for autism intervention. In *2015 37th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 3767–3770).
- Farwell, L. A., and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6), 510–523.
- Farwell, L. A., and Donchin, E. (1991). The truth will out: Interrogative polygraphy (“lie detection”) with event-related brain potentials. *Psychophysiology*, 28(5), 531–547.
- Fazli, S., Popescu, F., Danóczy, M., Blankertz, B., Müller, K.-R., and Grozea, C. (2009). Subject-independent mental state classification in single trials. *Neural networks*, 22(9), 1305–1312.
- Feng, L. (2010). *Automatic readability assessment*. City University of New York.
- Ferrez, P. W., and Millán, J. d. R. (2005). You are wrong!—automatic detection of interaction errors from brain waves. In *Proceedings of the 19th international joint conference on artificial intelligence*.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2), 179–188.
- Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3), 221.
- François, T., and Fairon, C. (2012). An “ai readability” formula for french as a foreign language. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 466–477).
- Friedman, D. B., and Hoffman-Goetz, L. (2006). A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education and Behavior*, 33(3), 352–373.
- Fumagalli, M., Giannicola, G., Rosa, M., Marceglia, S., Lucchiari, C., Mrakic-Spota, S., ... Priori, A. (2011). Conflict-dependent dynamic of subthalamic nucleus oscillations during moral decisions. *Social Neuroscience*, 6.
- Gannon, E., He, J., Gao, X., and Chaparro, B. (2016). Rsvp reading on a smart watch. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 1130–1134).
- Gateau, T., Ayaz, H., and Dehais, F. (2018). In silico vs. over the clouds: on-the-fly mental state estimation of aircraft pilots, using a functional near infrared spectroscopy based passive-bci. *Frontiers in human neuroscience*, 12, 187.

- Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., and Donchin, E. (1993). A neural system for error detection and compensation. *Psychological science*, 4(6), 385–390.
- George, L., and Lécuyer, A. (2010). An overview of research on " passive" brain-computer interfaces for implicit human-computer interaction. In *International conference on applied bionics and biomechanics icabb 2010-workshop w1 " brain-computer interfacing and virtual reality"*.
- Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., and Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in neuroscience*, 8, 385.
- Gevins, A., and Smith, M. E. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral cortex*, 10(9), 829–839.
- Gevins, A., Smith, M. E., McEvoy, L., and Yu, D. (1997). High-resolution eeg mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cerebral cortex (New York, NY: 1991)*, 7(4), 374–385.
- Ghazizadeh, M., Peng, Y., Lee, J. D., and Boyle, L. N. (2012). Augmenting the technology acceptance model with trust: Commercial drivers' attitudes towards monitoring and feedback..
- Goetz, J., and Kiesler, S. (2002). Cooperation with a robotic assistant..
- Goodfellow, A., and Bengio, Y. (2016). *Deep learning - ian goodfellow, yoshua bengio, aaron courville - google books*.
- Graesser, A. C., Louwerse, M. M., McNamara, D. S., Olney, A., Cai, Z., and Mitchell, H. H. (2007). Inference generation and cohesion in the construction of situation models: Some connections with computational linguistics. *Higher level language processes in the brain: Inference and comprehension processes*, 289–310.
- Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., and Pennebaker, J. (2014). Coh-matrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 115(2), 210–229.
- Grevet, E., Forge, K., Tadiello, S., Izac, M., Amadiou, E., Brunel, L., ... Jeunet-Kelway, C. (2023). Modeling the acceptability of bcis for motor rehabilitation after stroke: A large scale study on the general public. *Frontiers in Neuroergonomics*, 3, 1082901.
- Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., and Rao, R. P. (2008). Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 835–844).
- Gürkök, H., and Nijholt, A. (2012). Brain-computer interfaces for multimodal interaction: a survey and principles. *International Journal of Human-Computer Interaction*, 28(5), 292–307.

- Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of coling 2012* (pp. 1063–1080).
- Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, pp. 904–908).
- Hart, S. G., and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.
- Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87, 96–110.
- Hebert, R., Lehmann, D., Tan, G., Travis, F., and Arenander, A. (2005). Enhanced eeg alpha time-domain phase synchrony during transcendental meditation: Implications for cortical integration theory. *Signal Processing*, 85(11), 2213–2232.
- Hedin, B., and Lindgren, E. (2007). A comparison of presentation methods for reading on mobile phones. *IEEE Distributed Systems Online*, 8(6), 2–2.
- Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fnirs. *Frontiers in human neuroscience*, 7, 935.
- Hester, M., Werner, S., Greenwald, C., and Gunning, J. (2016). Exploring the effects of text length and difficulty on rsvp reading. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 1294–1298).
- Hettinger, L. J., Branco, P., Encarnacao, L. M., and Bonato, P. (2003). Neuroadaptive technologies: applying neuroergonomics to the design of advanced interfaces. *Theoretical Issues in Ergonomics Science*, 4(1-2), 220–237.
- Hinterberger, T., Neumann, N., Pham, M., Kübler, A., Grether, A., Hofmayer, N., ... Birbaumer, N. (2004). A multimodal brain-based feedback and communication system. *Experimental brain research*, 154(4), 521–526.
- Hitziger, S. (2015). *Modeling the variability of electrical activity in the brain* (Unpublished doctoral dissertation). Université Nice Sophia Antipolis.
- Holm, A., Lukander, K., Korpela, J., Sallinen, M., and Müller, K. M. (2009). Estimating brain load from the eeg. *TheScientificWorldJOURNAL*, 9, 639–651.
- Hudlicka, E., and Mcneese, M. D. (2002). Assessment of user affective and belief states for interface adaptation: Application to an air force pilot task. *User Modeling and User-Adapted Interaction*, 12(1), 1–47.
- IEEE Standards Association. (2020). *Standards roadmap: Neurotechnologies for brain-machine interfacing*.
- Ienca, M., Haselager, P., and Emanuel, E. J. (2018). Brain leaks and consumer neurotechnology. *Nature biotechnology*, 36(9), 805–810.

- Iversen, J. R., and Makeig, S. (2019). Meg/eeg data analysis using eeglab. *Magnetoencephalography: From Signals to Dynamic Cortical Networks*, 391–406.
- Jasper, H. H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.*, 10, 370–375.
- Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., and Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1), 20–31.
- Jung, T.-P., Makeig, S., McKeown, M. J., Bell, A. J., Lee, T.-W., and Sejnowski, T. J. (2001). Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE*, 89(7), 1107–1122.
- Just, M. A., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1), 122.
- Kaimara, P., Plerou, A., and Deliyannis, I. (2020). Cognitive enhancement and brain-computer interfaces: potential boundaries and risks. In *Genedis 2018* (pp. 275–283). Springer.
- Karray, F., Karray, F. O., and Silva, C. W. D. (2004). *Soft computing and intelligent systems design: theory, tools, and applications*. Pearson Education.
- Kendeou, P., McMaster, K. L., and Christ, T. J. (2016). Reading comprehension: Core components and processes. *Policy Insights from the Behavioral and Brain Sciences*, 3(1), 62–69.
- Khaliliardali, Z., Chavarriaga, R., Gheorghe, L. A., and Millán, J. d. R. (2012). Detection of anticipatory brain potentials during car driving. In *2012 annual international conference of the IEEE engineering in medicine and biology society* (pp. 3829–3832).
- Kim, K.-T., Suk, H.-I., and Lee, S.-W. (2016). Commanding a brain-controlled wheelchair using steady-state somatosensory evoked potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(3), 654–665.
- Kim, S. K., and Kirchner, E. A. (2013). Classifier transferability in the detection of error related potentials from observation to interaction. In *2013 IEEE international conference on systems, man, and cybernetics* (pp. 3360–3365).
- Kim, S. K., Kirchner, E. A., Stefes, A., and Kirchner, F. (2017). Intrinsic interactive reinforcement learning—using error-related potentials for real world human-robot interaction. *Scientific reports*, 7(1), 1–16.
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel* (Tech. Rep.). Naval Technical Training Command Millington TN Research Branch.
- Kindermans, P.-J., Schreuder, M., Schrauwen, B., Müller, K.-R., and Tangermann, M. (2014). True zero-training brain-computer interfacing—an online study. *PLoS one*, 9(7), e102504.

- Kintsch, W., and Walter Kintsch, C. B. E. M. A. F. R. S. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.
- Kleinberger, T., Jedlitschka, A., Storf, H., Steinbach-Nordmann, S., and Prueckner, S. (2009). An approach to and evaluations of assisted living systems using ambient intelligence for emergency monitoring and prevention. In (Vol. 5615 LNCS).
- Kleiner, M., Brainard, D., and Pelli, D. (2007). What's new in psychtoolbox-3?
- Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Müller, K.-R., Curio, G., ... others (2007). Improving human performance in a real operating environment through real-time mental workload detection. *Toward brain-computer interfacing*, 409422, 409–422.
- Kotchetkov, I. S., Hwang, B. Y., Appelboom, G., Kellner, C. P., and Connolly, E. S. (2010). Brain-computer interfaces: military, neurosurgical, and ethical perspective. *Neurosurgical focus*, 28(5), E25.
- Kothe, C. A., and Makeig, S. (2013). Bcilib: a platform for brain–computer interface development. *Journal of neural engineering*, 10(5), 056014.
- Kotsiantis, S. B. (2007). *Supervised machine learning: A review of classification techniques* (Vol. 31).
- Kramer, A. F. (2020). Physiological metrics of mental workload: A review of recent progress. *Multiple-task performance*, 279–328.
- Krauledat, J. M. (2008). Analysis of nonstationarities in eeg signals for improving brain-computer interface performance.
- Krauledat, M., Tangermann, M., Blankertz, B., and Müller, K.-R. (2008). Towards zero training for brain-computer interfacing. *PloS one*, 3(8), e2967.
- Krause, C. M., Pesonen, M., and Hämäläinen, H. (2010). Brain oscillatory 4–30 hz electroencephalogram responses in adolescents during a visual memory task. *Neuroreport*, 21(11), 767–771.
- Krause, C. M., Sillanmäki, L., Koivisto, M., Saarela, C., Häggqvist, A., Laine, M., and Hämäläinen, H. (2000). The effects of memory load on event-related eeg desynchronization and synchronization. *Clinical neurophysiology*, 111(11), 2071–2078.
- Kreilinger, A., Neuper, C., and Müller-Putz, G. R. (2012). Error potential detection during continuous movement of an artificial arm controlled by brain–computer interface. *Medical and biological engineering and computing*, 50(3), 223–230.
- Krol, L. R., Andreessen, L. M., Podgorska, A., Makarov, N., and Zander, T. O. (2018). Passive brain-computerinterfacinginthemuseumofstillness. *Proceedings of theArtistic BCIWorkshopattheSIGCHIConferenceonHumanFactorsinComputingSystems (CHI)*.

- Krol, L. R., Freytag, S.-C., Fleck, M., Gramann, K., and Zander, T. O. (2016). A task-independent workload classifier for neuroadaptive technology: Preliminary data. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 003171–003174).
- Krol, L. R., Mousavi, M., De Sa, V., and Zander, T. (2018). Towards classifier visualisation in 3d source space. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 71–76).
- Krol, L. R., and Zander, T. O. (2017). Passive bci-based neuroadaptive systems. In *Gbcic*.
- Krusienski, D. J., Sellers, E. W., McFarland, D. J., Vaughan, T. M., and Wolpaw, J. R. (2008). Toward enhanced p300 speller performance. *Journal of neuroscience methods*, 167(1), 15–21.
- Kübler, A., Holz, E. M., Riccio, A., Zickler, C., Kaufmann, T., Kleih, S. C., ... Mattia, D. (2014). The user-centered design as novel perspective for evaluating the usability of bci-controlled applications. *PloS one*, 9(12), e112392.
- Kübler, A., Nijboer, F., Mellinger, J., Vaughan, T. M., Pawelzik, H., Schalk, G., ... Wolpaw, J. R. (2005). Patients with als can use sensorimotor rhythms to operate a brain-computer interface. *Neurology*, 64(10), 1775–1777.
- Kuhn, M. R., Schwanenflugel, P. J., and Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading research quarterly*, 45(2), 230–251.
- Kumar, A., Gao, L., Pirogova, E., and Fang, Q. (2019). A review of error-related potential-based brain-computer interfaces for motor impaired people. *IEEE Access*, 7, 142451–142466.
- Lahat, A., Helwig, C. C., and Zelazo, P. D. (2013). An event-related potential study of adolescents' and young adults' judgments of moral and social conventional violations. *Child Development*, 84.
- Lakshmi, M. R., Prasad, T., and Prakash, D. V. C. (2014). Survey on eeg signal processing methods. *International journal of advanced research in computer science and software engineering*, 4(1).
- Lample, G., and Chaplot, D. S. (2017). Playing fps games with deep reinforcement learning..
- Lang, P. J., Bradley, M. M., Cuthbert, B. N., et al. (1997). International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1(39-58), 3.
- Leuthold, H., Kunkel, A., Mackenzie, I. G., and Filik, R. (2014). Online processing of moral transgressions: Erp evidence for spontaneous evaluation. *Social Cognitive and Affective Neuroscience*, 10.
- Lew, G. S., and Polich, J. (1993). P300, habituation, and response mode. *Physiology and behavior*, 53(1), 111–117.

- Li, Q., Ding, D., and Conti, M. (2015). Brain-computer interface applications: Security and privacy challenges. In *2015 IEEE conference on communications and network security (CNS)* (pp. 663–666).
- Lim, C. G., Lee, T. S., Guan, C., Fung, D. S. S., Zhao, Y., Teng, S. S. W., ... Krishnan, K. R. R. (2012). A brain-computer interface based attention training program for treating attention deficit hyperactivity disorder. *PloS one*, 7(10), e46692.
- Lin, C.-T., Chang, C.-J., Lin, B.-S., Hung, S.-H., Chao, C.-F., and Wang, I.-J. (2010). A real-time wireless brain-computer interface system for drowsiness detection. *IEEE transactions on biomedical circuits and systems*, 4(4), 214–222.
- Loke, I. C., Evans, A. D., and Lee, K. (2011). The neural correlates of reasoning about prosocial-helping decisions: An event-related brain potentials study. *Brain Research*, 1369.
- Lotte, F. (2008). *Study of electroencephalographic signal processing and classification techniques towards the use of brain-computer interfaces in virtual reality applications* (Unpublished doctoral dissertation). INSA de Rennes.
- Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., and Yger, F. (2018). A review of classification algorithms for eeg-based brain-computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3), 031005.
- Lu, S., Guan, C., and Zhang, H. (2009). Unsupervised brain computer interface based on intersubject information and online adaptation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 17(2), 135–145.
- Luck, S. J. (2012). Event-related potentials.
- Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.
- Luck, S. J., Woodman, G. F., and Vogel, E. K. (2000). Event-related potential studies of attention. *Trends in cognitive sciences*, 4(11), 432–440.
- Ludwig, K. A., Miriani, R. M., Langhals, N. B., Joseph, M. D., Anderson, D. J., and Kipke, D. R. (2009). Using a common average reference to improve cortical neuron recordings from microelectrode arrays. *Journal of neurophysiology*, 101(3), 1679–1689.
- Makeig, S., Bell, A., Jung, T.-P., and Sejnowski, T. J. (1995). Independent component analysis of electroencephalographic data. *Advances in neural information processing systems*, 8.
- Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1), 177–190.
- Mathewson, K. E., Harrison, T. J., and Kizuk, S. A. (2017). High and dry? comparing active dry eeg electrodes to active and passive wet electrodes. *Psychophysiology*, 54(1), 74–82.

- Mayeli, A., Zotev, V., Refai, H., and Bodurka, J. (2016). Real-time eeg artifact correction during fmri using ica. *Journal of neuroscience methods*, 274, 27–37.
- McCullagh, P., Lightbody, G., Zygierewicz, J., and Kernohan, W. G. (2014). Ethical challenges associated with the development and deployment of brain computer interface technology. *Neuroethics*, 7(2), 109–122.
- McFarland, D. J., Lefkowitz, A. T., and Wolpaw, J. R. (1997). Design and operation of an eeg-based brain-computer interface with digital signal processing technology. *Behavior Research Methods, Instruments, and Computers*, 29(3), 337–345.
- McNamara, D. S., Kintsch, E., Songer, N. B., and Kintsch, W. (1996). Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction*, 14(1), 1–43.
- Miltner, W. H., Braun, C. H., and Coles, M. G. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a “generic” neural system for error detection. *Journal of cognitive neuroscience*, 9(6), 788–798.
- Minguillon, J., Lopez-Gordo, M. A., and Pelayo, F. (2017). Trends in eeg-bci for daily-life: Requirements for artifact removal. *Biomedical Signal Processing and Control*, 31, 407–418.
- Miranda, E. R., and Castet, J. (2014). *Guide to brain-computer music interfacing*. Springer.
- Misselhorn, C. (2018). Artificial morality. concepts, issues and challenges. *Society*, 55.
- Misselhorn, C. (2020). Artificial systems with moral capacities? a research design and its implementation in a geriatric care system. *Artificial Intelligence*, 278, 103179.
- Misselhorn, C., Pompe, U., and Stapleton, M. (2013). Ethical considerations regarding the use of social robots in the fourth age. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 26.
- Miyakoshi, M. (2017). *fullRankAveRef*. Retrieved 2021-07-11, from http://sccn.ucsd.edu/wiki/Plugin_list_process
- Moravcsik, J. E., and Kintsch, W. (1993). Writing quality, reading skills, and domain knowledge as factors in text comprehension. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 47(2), 360.
- Mühl, C., Gürkök, H., Plass-Oude Bos, D., Thurlings, M. E., Scherffig, L., Duvinage, M., ... Heylen, D. (2010). Bacteria hunt. *Journal on multimodal user interfaces*, 4(1), 11–25.
- Müller, K.-R., Anderson, C. W., and Birch, G. E. (2003). Linear and nonlinear methods for brain-computer interfaces. *IEEE transactions on neural systems and rehabilitation engineering*, 11(2), 165–169.

- Müller-Putz, G., Scherer, R., Brunner, C., Leeb, R., and Pfurtscheller, G. (2008). Better than random: a closer look on bci results. *International journal of bioelectromagnetism*, 10(ARTICLE), 52–55.
- Müller-Putz, G., Scherer, R., Pfurtscheller, G., and Rupp, R. (2005). Eeg-based neuroprosthesis control: a step towards clinical practice. *Neuroscience letters*, 382(1-2), 169–174.
- Müller, K.-R., Krauledat, M., Dornhege, G., Curio, G., and Blankertz, B. (2004). Machine learning techniques for brain-computer interfaces. *Machine Learning*, 49.
- Müller, V. C., and Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. *Fundamental Issues of Artificial Intelligence*.
- Nakanishi, I., Baba, S., Ozaki, K., and Li, S. (2013). Using brain waves as transparent biometrics for on-demand driver authentication. *International journal of biometrics*, 5(3-4), 288–305.
- Nasehi, S., Pourghassem, H., and Isfahan, I. (2012). An optimal eeg-based emotion recognition algorithm using gabor. *WSEAS transactions on signal processing*, 3(8), 87–99.
- Newborn, M. (2012). *Kasparov versus deep blue: Computer chess comes of age*. Springer Science and Business Media.
- Nijboer, F., Van De Laar, B., Gerritsen, S., Nijholt, A., and Poel, M. (2015). Usability of three electroencephalogram headsets for brain-computer interfaces: a within subject comparison. *Interacting with computers*, 27(5), 500–511.
- Nijholt, A. (2019). *Brain art: Brain-computer interfaces for artistic expression*. Springer.
- Nijholt, A., Bos, D. P.-O., and Reuderink, B. (2009). Turning shortcomings into challenges: Brain-computer interfaces for games. *Entertainment computing*, 1(2), 85–94.
- Nowlis, D. P., and Kamiya, J. (1970). The control of electroencephalographic alpha rhythms through auditory feedback and the associated mental activity. *Psychophysiology*, 6(4), 476–484.
- Ojala, M., and Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of machine learning research*, 11(6).
- Oostenveld, R., and Praamstra, P. (2001). The five percent electrode system for high-resolution eeg and erp measurements. *Clinical neurophysiology*, 112(4), 713–719.
- Öquist, G., and Goldstein, M. (2003). Towards an improved readability on mobile devices: evaluating adaptive rapid serial visual presentation. *Interacting with Computers*, 15(4), 539–558.

- Osselton, J. (1965). Acquisition of eeg data by bipolar unipolar and average reference methods: a theoretical comparison. *Electroencephalography and clinical neurophysiology*, 19(5), 527–528.
- Oudeyer, P.-Y., Gottlieb, J., and Lopes, M. (2016). Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. *Progress in brain research*, 229, 257–284.
- Palumbo, A., Amato, F., Calabrese, B., Cannataro, M., Cocorullo, G., Gambardella, A., ... others (2010). An embedded system for eeg acquisition and processing for brain computer interface applications. In *Wearable and autonomous biomedical devices and systems for smart environment* (pp. 137–154). Springer.
- Parasuraman, R., and Rizzo, M. (2007). Introduction to neuroergonomics. *Neuroergonomics: The brain at work*, 3–12.
- Parra, L. C., Spence, C. D., Gerson, A. D., and Sajda, P. (2003). Response error correction—a demonstration of improved human-machine performance using real-time eeg monitoring. *IEEE transactions on neural systems and rehabilitation engineering*, 11(2), 173–177.
- Pasqualotto, E., Matuz, T., Federici, S., Ruf, C. A., Bartl, M., Olivetti Belardinelli, M., ... Halder, S. (2015). Usability and workload of access technology for people with severe motor impairment: a comparison of brain-computer interfacing and eye tracking. *Neurorehabilitation and neural repair*, 29(10), 950–957.
- Pawar, D., and Dhage, S. (2020). Feature extraction methods for electroencephalography based brain-computer interface: A review. *IAENG International Journal of Computer Science*, 47(3).
- Pawlitzki, J., Klaproth, O., Krol, L. R., and Zander, T. O. (2021). Automation surprise in the neuroadaptive cockpit.
- Perfetti, C. A. (1985). *Reading ability*. oxford university Press.
- Pesonen, M., Hämäläinen, H., and Krause, C. M. (2007). Brain oscillatory 4–30 hz responses during a visual n-back memory task with varying memory load. *Brain research*, 1138, 171–177.
- Petersen, S. E., and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer speech and language*, 23(1), 89–106.
- Pfurtscheller, G., Allison, B. Z., Bauernfeind, G., Brunner, C., Solis Escalante, T., Scherer, R., ... Birbaumer, N. (2010). The hybrid bci. *Frontiers in neuroscience*, 4, 3.
- Pfurtscheller, G., Brunner, C., Schlögl, A., and Da Silva, F. L. (2006). Mu rhythm (de) synchronization and eeg single-trial classification of different motor imagery tasks. *NeuroImage*, 31(1), 153–159.
- Pfurtscheller, G., and Da Silva, F. L. (1999). Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11), 1842–1857.

- Picard, R. W. (2000). *Affective computing*. MIT press.
- Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S., Johnson, R., ... others (2000). Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology*, 37(2), 127–152.
- Pion-Tonachini, L., Kreutz-Delgado, K., and Makeig, S. (2019). Iclabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, 198, 181-197.
- Polich, J. (2007). Updating p300: an integrative theory of p3a and p3b. *Clinical neurophysiology*, 118(10), 2128–2148.
- Pritchard, W. S. (1981). Psychophysiology of p300. *Psychological bulletin*, 89(3), 506.
- Protzak, J., Ihme, K., and Zander, T. O. (2013). A passive brain-computer interface for supporting gaze-based human-machine interaction. In *International conference on universal access in human-computer interaction* (pp. 662–671).
- Prpa, M., and Pasquier, P. (2019). Brain-computer interfaces in contemporary art: a state of the art and taxonomy. *Brain art*, 65–115.
- Radüntz, T., Meffert, B., et al. (2019). User experience of 7 mobile electroencephalography devices: comparative study. *JMIR mHealth and uHealth*, 7(9), e14474.
- Rak, R. J., Kołodziej, M., and Majkowski, A. (2012). Brain-computer interface as measurement and control system the review paper. *Metrology and Measurement Systems*, 427–444.
- Ramadan, R. A., and Vasilakos, A. V. (2017). Brain computer interface: control signals review. *Neurocomputing*, 223, 26–44.
- Rao, T. K., Lakshmi, M. R., and Prasad, T. (2012). An exploration on brain computer interface and its recent trends. *arXiv preprint arXiv:1211.2737*.
- Rashid, M., Sulaiman, N., PP Abdul Majeed, A., Musa, R. M., Bari, B. S., Khatun, S., et al. (2020). Current status, challenges, and possible solutions of eeg-based brain-computer interface: a comprehensive review. *Frontiers in neurorobotics*, 25.
- Riedl, M. O., and Harrison, B. (2016). Using stories to teach human values to artificial agents. In (Vol. WS-16-01 - WS-16-15).
- Roy, R. N., Charbonnier, S., Campagne, A., and Bonnet, S. (2016). Efficient mental workload estimation using task-independent eeg features. *Journal of neural engineering*, 13(2), 026019.
- Roy, R. N., Drougard, N., Gateau, T., Dehais, F., and Chanel, C. P. (2020). How can physiological computing benefit human-robot interaction? *Robotics*, 9(4), 100.
- Russell, S. (2016). *Should we fear supersmart robots?* (Vol. 314).

- Russell, S., and Norvig, P. (2010). *Artificial intelligence a modern approach third edition*.
- Saab, J., Battes, B., Grosse-Wentrup, M., Scherer, R., Billinger, M., Kreilinger, A., et al. (2011). *Simultaneous eeg recordings with dry and wet electrodes in motor-imagery*. Citeseer.
- Saffiotti, A., and Broxvall, M. (2005). Peis ecologies: Ambient intelligence meets autonomous robotics. In (Vol. 121).
- Saha, S., Mamun, K. A., Ahmed, K., Mostafa, R., Naik, G. R., Khandoker, A., ... Baumert, M. (2019). Progress in brain computer interfaces: challenges and trends. *arXiv preprint arXiv:1901.03442*.
- Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of educational psychology*, 76(4), 647.
- Sarma, G. P., and Hay, N. J. (2017). Mammalian value systems. *Informatica (Slovenia)*, 41.
- Sarter, N. B., Woods, D. D., Billings, C. E., et al. (1997). Automation surprises. *Handbook of human factors and ergonomics*, 2, 1926–1943.
- Scharinger, C., Kammerer, Y., and Gerjets, P. (2015). Pupil dilation and eeg alpha frequency band power reveal load on executive functions for link-selection processes during text reading. *PloS one*, 10(6), e0130608.
- Scharinger, C., Soutschek, A., Schubert, T., and Gerjets, P. (2017). Comparison of the working memory load in n-back and working memory span tasks by means of eeg frequency band power and p300 amplitude. *Frontiers in human neuroscience*, 11, 6.
- Scheffers, M. K., and Coles, M. G. (2000). Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1), 141.
- Schiliro, F., Moustafa, N., and Beheshti, A. (2020). Cognitive privacy: Ai-enabled privacy using eeg signals in the internet of things. In *2020 ieee 6th international conference on dependability in sensor, cloud and big data systems and application (dependsys)* (pp. 73–79).
- Schmidt, N. M., Blankertz, B., and Treder, M. S. (2012). Online detection of error-related potentials boosts the performance of mental typewriters. *BMC neuroscience*, 13(1), 1–13.
- Schultheis, H., and Jameson, A. (2004). Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In *International conference on adaptive hypermedia and adaptive web-based systems* (pp. 225–234).

- Schultze-Kraft, M., Dähne, S., Gugler, M., Curio, G., and Blankertz, B. (2016). Unsupervised classification of operator workload from brain signals. *Journal of neural engineering*, 13(3), 036008.
- Schwarm, S. E., and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)* (pp. 523–530).
- Segalowitz, S. J., and Barnes, K. L. (1993). The reliability of erp components in the auditory oddball paradigm. *Psychophysiology*, 30(5), 451–459.
- Sellers, E. W., and Donchin, E. (2006). A p300-based brain-computer interface: initial tests by als patients. *Clinical neurophysiology*, 117(3), 538–548.
- Shanahan, M. (2000). M. shanahan, solving the frame problem. *Artificial Intelligence*, 123.
- Shanahan, M. (2019). *The technological singularity*.
- Sharma, R., Hicks, S., Berna, C. M., Kennard, C., Talbot, K., and Turner, M. R. (2011). Oculomotor dysfunction in amyotrophic lateral sclerosis: a comprehensive review. *Archives of Neurology*, 68(7), 857–861.
- Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P., and Müller, K.-R. (2006). Towards adaptive classification for bci. *Journal of neural engineering*, 3(1), R13.
- Shih, J. J., Krusienski, D. J., and Wolpaw, J. R. (2012). Brain-computer interfaces in medicine. In *Mayo clinic proceedings* (Vol. 87, pp. 268–279).
- Shriram, R., Sundhararajan, M., and Daimiwal, N. (2013). Eeg based cognitive workload assessment for maximum efficiency. *Int. Organ. Sci. Res. IOSR*, 7, 34–38.
- Si, L., and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the tenth international conference on information and knowledge management* (pp. 574–576).
- Soares, N., and Fallenstein, B. (2014). Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report*, 8.
- Spüler, M., Bensch, M., Kleih, S., Rosenstiel, W., Bogdan, M., and Kübler, A. (2012). Online use of error-related potentials in healthy users and people with severe motor impairment increases performance of a p300-bci. *Clinical Neurophysiology*, 123(7), 1328–1337.
- Spüler, M., Krumpe, T., Walter, C., Scharinger, C., Rosenstiel, W., and Gerjets, P. (2017). Brain-computer interfaces for educational applications. In *Informational environments* (pp. 177–201). Springer.
- Spüler, M., and Niethammer, C. (2015). Error-related potentials during continuous feedback: using eeg to detect errors of different type and severity. *Frontiers in human neuroscience*, 9, 155.

- Sweller, J., Van Merriënboer, J. J., and Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational psychology review*, 10(3), 251–296.
- Tandle, A., Jog, N., D’cunha, P., and Chheta, M. (2015). Classification of artefacts in eeg signal recordings and overview of removing techniques. *International Journal of Computer Applications*, 975, 8887.
- Tapia, D. I., Abraham, A., Corchado, J. M., and Alonso, R. S. (2010). Agents and ambient intelligence: Case studies. *Journal of Ambient Intelligence and Humanized Computing*, 1.
- Tapus, A., and Matarić, M. J. (2008). User personality matching with a hands-off robot for post-stroke rehabilitation therapy. In (Vol. 39).
- Tautan, A.-M., Mihajlovic, V., Chen, Y.-H., Grundlehner, B., Penders, J., and Serdijn, W. A. (2014). Signal quality in dry electrode eeg and the relation to skin-electrode contact impedance magnitude. In *Biodevices* (pp. 12–22).
- Taylor, S. F., Stern, E. R., and Gehring, W. J. (2007). Neural systems for error monitoring: recent findings and theoretical perspectives. *The Neuroscientist*, 13(2), 160–172.
- Teplan, M., et al. (2002). Fundamentals of eeg measurement. *Measurement science review*, 2(2), 1–11.
- Thomas, K. P., and Vinod, A. (2017). Toward eeg-based biometric systems: The great potential of brain-wave-based biometrics. *IEEE Systems, Man, and Cybernetics Magazine*, 3(4), 6–15.
- Thongpang, S., Richner, T. J., Brodnick, S. K., Schendel, A., Kim, J., Wilson, J. A., ... others (2011). A micro-electrocorticography platform and deployment strategies for chronic bci applications. *Clinical EEG and neuroscience*, 42(4), 259–265.
- Tibus, M., Heier, A., and Schwan, S. (2013). Do films make you learn? inference processes in expository film comprehension. *Journal of Educational Psychology*, 105(2).
- Tsui, C. S. L., Gan, J. Q., and Hu, H. (2011). A self-paced motor imagery based brain-computer interface for robotic wheelchair control. *Clinical EEG and neuroscience*, 42(4), 225–229.
- Turing, A. M., and Haugeland, J. (1950). *Computing machinery and intelligence*. MIT Press Cambridge, MA.
- Vajjala, S., and Meurers, D. (2014). Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2), 194–222.
- Vajjala Balakrishna, S. (2015). *Analyzing text complexity and text simplification: Connecting linguistics, processing and educational applications* (Unpublished doctoral dissertation). Universitat Tübingen.

- Val-Calvo, M., Álvarez-Sánchez, J. R., Ferrández-Vicente, J. M., and Fernández, E. (2019). Optimization of real-time eeg artifact removal and emotion estimation for human-robot interaction applications. *Frontiers in Computational Neuroscience*, 13, 80.
- Van Erp, J., Lotte, F., and Tangermann, M. (2012). Brain-computer interfaces: beyond medical applications. *Computer*, 45(4), 26–34.
- Vapnik, V. (1999a). *The nature of statistical learning theory*. Springer science and business media.
- Vapnik, V. (1999b). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5), 988–999.
- Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science and business media.
- Vařeka, L., and Ladouce, S. (2021). Prediction of navigational decisions in the real-world: a visual p300 event-related potentials brain-computer interface. *International Journal of Human–Computer Interaction*, 37(14), 1375–1389.
- Vecchiato, G., Babiloni, F., Astolfi, L., Toppi, J., Cherubino, P., Dai, J., ... Wei, D. (2011). Enhance of theta eeg spectral activity related to the memorization of commercial advertisings in chinese and italian subjects. In *2011 4th international conference on biomedical engineering and informatics (bmei)* (Vol. 3, pp. 1491–1494).
- Venkatesan, R., and Er, M. J. (2016). A novel progressive learning technique for multi-class classification. *Neurocomputing*, 207, 310–321.
- Vidal, J. J. (1973). Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering*, 2(1), 157–180.
- Vygotsky, L. S., and Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.
- Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology*, 12.
- Wallach, W., and Allen, C. (2009). *Moral machines: Teaching robots right from wrong*.
- Wallach, W., Allen, C., and Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI and Society*, 22.
- Wallot, S., O'Brien, B. A., Haussmann, A., Kloos, H., and Lyby, M. S. (2014). The role of reading time complexity and reading speed in text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6), 1745.
- Walter, C., Rosenstiel, W., Bogdan, M., Gerjets, P., and Spüler, M. (2017). Online eeg-based workload adaptation of an arithmetic learning environment. *Frontiers in human neuroscience*, 11, 286.

- Watkins, C. J., and Dayan, P. (1992). Technical note: Q-learning. *Machine Learning*, 8.
- Wickens, C. D. (2008). Multiple resources and mental workload. *Human factors*, 50(3), 449–455.
- Wirth, C., Dockree, P., Harty, S., Lacey, E., and Arvaneh, M. (2019). Towards error categorisation in bci: single-trial eeg classification between different errors. *Journal of neural engineering*, 17(1), 016008.
- Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6), 767–791.
- Wolpaw, J. R., and Wolpaw, E. W. (2012). Brain-computer interfaces: something new under the sun. *Brain-computer interfaces: principles and practice*, 14.
- Yagi, Y., Coburn, K. L., Estes, K. M., and Arruda, J. E. (1999). Effects of aerobic exercise and gender on visual and auditory p300, reaction time, and accuracy. *European journal of applied physiology and occupational physiology*, 80(5), 402–408.
- Yampolskiy, R. V. (2019). Personal universes: A solution to the multi-agent value alignment problem. *arXiv preprint arXiv:1901.01851*.
- Yang, M. H., Kriegman, D. J., and Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24.
- Yoder, K. J., and Decety, J. (2014). Spatiotemporal neural dynamics of moral judgment: A high-density erp study. *Neuropsychologia*, 60.
- Yousefi, R., Sereshkeh, A. R., and Chau, T. (2018). Exploiting error-related potentials in cognitive task based bci. *Biomedical Physics and Engineering Express*, 5(1), 015023.
- Yuksel, B. F., Oleson, K. B., Harrison, L., Peck, E. M., Afergan, D., Chang, R., and Jacob, R. J. (2016). Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 5372–5384).
- Zander, T. O. (2012). Utilizing brain-computer interfaces for human-machine systems.
- Zander, T. O., Andreessen, L. M., Berg, A., Bleuel, M., Pawlitzki, J., Zawallich, L., ... Gramann, K. (2017). Evaluation of a dry eeg system for application of passive brain-computer interfaces in autonomous driving. *Frontiers in human neuroscience*, 11, 78.
- Zander, T. O., Brönstrup, J., Lorenz, R., and Krol, L. R. (2014). Towards bci-based implicit control in human–computer interaction. *Advances in Physiological Computing*, 67–90.

- Zander, T. O., and Jatzev, S. (2011). Context-aware brain–computer interfaces: exploring the information space of user, technical system and environment. *Journal of Neural Engineering*, 9(1), 016003.
- Zander, T. O., and Kothe, C. (2011). Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *Journal of neural engineering*, 8(2), 025005.
- Zander, T. O., Kothe, C., Jatzev, S., and Gaertner, M. (2010). Enhancing human–computer interaction with input from active and passive brain–computer interfaces. In *Brain-computer interfaces* (pp. 181–199). Springer.
- Zander, T. O., Kothe, C., Welke, S., and Rötting, M. (2009). Utilizing secondary input from passive brain–computer interfaces for enhancing human–machine interaction. In *International conference on foundations of augmented cognition* (pp. 759–771).
- Zander, T. O., Krol, L. R., Birbaumer, N. P., and Gramann, K. (2016). Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity. *Proceedings of the National Academy of Sciences*, 113(52), 14898–14903.
- Zander, T. O., Lehne, M., Ihme, K., Jatzev, S., Correia, J., Kothe, C., ... Nijboer, F. (2011). A dry eeg-system for scientific research and brain–computer interfaces. *Frontiers in neuroscience*, 5, 53.
- Zhang, X., Krol, L. R., and Zander, T. O. (2018). Towards task-independent workload classification: shifting from binary to continuous classification. In *2018 IEEE international conference on systems, man, and cybernetics (smc)* (pp. 556–561).
- Zhu, D., Bieger, J., Garcia Molina, G., and Aarts, R. M. (2010). A survey of stimulation methods used in ssvp-based bcis. *Computational intelligence and neuroscience*, 2010.
- Zotев, V., Phillips, R., Yuan, H., Misaki, M., and Bodurka, J. (2014). Self-regulation of human brain activity using simultaneous real-time fmri and eeg neurofeedback. *NeuroImage*, 85, 985–995.
- Zschocke, S., and Hansen, H.-C. (2011). *Klinische elektroenzephalographie*. Springer-Verlag.

List of Figures

2.1	Overview of the components of a BCI	7
2.2	Overview of different invasive and non-invasive neurophysiological measurement techniques	9
2.3	International 10/20 system for electrode placement	10
2.4	Typical ERP components in the EEG	12
2.5	Oscillations of the frequency bands in the EEG	13
2.6	Overview of the independent component analysis (ICA) model	16
2.7	Typical setup of a BCI experiment in the laboratory	23
3.1	The RLR-game experimental paradigm	56
3.2	Experimental setup for playing the RLR-game	56
3.3	Errors occurring at a red letter in the RLR-game, when 90 degree rotation is expected	57
3.4	Errors occurring at a green letter in the RLR-game, when 30 degree rotation is expected	57
3.5	Grand average difference curves of the ERP (error minus no error) of the first recording session at electrode sites Fz, Cz and Pz with permutation test results	60
3.6	Grand average ERPs of the first recording session at electrode sites Fz, Cz and Pz	61
3.7	Grand average patterns of classifier models from the first recording session (includes eye activity)	62
3.8	Weighted dipole density plots of the subject-specific classifier including data from eye components in time window 150-200 ms from the first recording session	63
3.9	Weighted dipole density plots of the subject-specific classifier including data from eye components in time window 200-250 ms from the first recording session	64
3.10	Weighted dipole density plots of the subject-specific classifier including data from eye components in time window 250-300 ms from the first recording session	64
3.11	Weighted dipole density plots of the subject-specific classifier including data from eye components in time window 300-350 ms from the first recording session	65
3.12	Weighted dipole density plots of the subject-specific classifier including data from eye components in time window 350-400 ms from the first recording session	65
3.13	Weighted dipole density plots of the subject-specific classifier including data from eye components in time window 400-450 ms from the first recording session	66

3.14	Weighted dipole density plots of the subject-specific classifier including data from eye components in time window 450-500 ms from the first recording session	66
3.15	Grand average difference curve of the ERP (error minus no error) data cleaned from eye activity through ICA of the first recording session at electrode sites Fz, Cz and Pz with permutation test results . . .	72
3.16	Grand average ERPs of data cleaned from eye activity through ICA at electrode sites Fz, Cz and Pz from the first recording session	73
3.17	Grand average patterns of classifier models from the first recording session (cleaned from eye activity)	75
3.18	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 150-200 ms from the first recording session	75
3.19	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 200-250 ms from the first recording session	76
3.20	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 250-300 ms from the first recording session	76
3.21	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 300-350 ms from the first recording session	77
3.22	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 350-400 ms from the first recording session	77
3.23	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 400-450 ms from the first recording session	78
3.24	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 450-500 ms from the first recording session	78
3.25	Patterns of the group classifier model trained on data from the first recording session cleaned from eye activity	85
3.26	Weighted dipole density plots of the group classifier cleaned from eye components in time window 150-200 ms from the first recording session	86
3.27	Weighted dipole density plots of the group classifier cleaned from eye components in time window 200-250 ms from the first recording session	87
3.28	Weighted dipole density plots of the group classifier cleaned from eye components in time window 250-300 ms from the first recording session	87
3.29	Weighted dipole density plots of the group classifier cleaned from eye components in time window 300-350 ms from the first recording session	88
3.30	Weighted dipole density plots of the group classifier cleaned from eye components in time window 350-400 ms from the first recording session	88

3.31	Weighted dipole density plots of the group classifier cleaned from eye components in time window 400-450 ms from the first recording session	89
3.32	Weighted dipole density plots of the group classifier cleaned from eye components in time window 450-500 ms from the first recording session	89
3.33	The three different electrode setups: all 50 electrodes from recording sessions 1 and 2, the symmetrically reduced electrode setup with 32 electrodes and the 24 electrode setup with the final omission of frontal electrodes.	98
3.34	Patterns of subject-specific classifiers trained on the 50 electrode setup with all data from recording sessions 1 and 2	99
3.35	Patterns of subject-specific classifiers trained on the 32 electrode setup with all data from recording sessions 1 and 2	100
3.36	Patterns of subject-specific classifiers trained on the 24 electrode setup with all data from recording sessions 1 and 2	100
3.37	Patterns of the group classifier trained on the 50 electrode setup with all data from recording sessions 1 and 2	102
3.38	Patterns of the group classifier trained on the 32 electrode setup with all data from recording sessions 1 and 2	103
3.39	Patterns of the group classifier trained on the 24 electrode setup with all data from recording sessions 1 and 2	103
3.40	Grand average ERPs of the third recording session at electrode sites Fz, Cz and Pz (cleaned from eye activity)	110
3.41	Grand average difference curve of the ERP (error minus no error) data cleaned from eye activity through ICA of the third recording session at electrode sites Fz, Cz and Pz with permutation test results . . .	111
3.42	Grand average patterns of subject-specific classifier models from the third recording session (cleaned from eye activity)	112
3.43	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 150-200 ms from the third recording session	112
3.44	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 200-250 ms from the third recording session	113
3.45	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 250-300 ms from the third recording session	113
3.46	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 300-350 ms from the third recording session	114
3.47	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 350-400 ms from the third recording session	114
3.48	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 400-450 ms from the third recording session	115

3.49	Weighted dipole density plots of the subject-specific classifier cleaned from eye components in time window 450-500 ms from the third recording session	115
4.1	Screenshots of the mental workload paradigm and the text reading paradigm	135
4.2	Experimental procedure including the mental workload paradigm and the text reading paradigm	138
4.3	Grand average spatial patterns available to the mental workload task-independent predictive model	145
4.4	Subject-wise predictive values of mental workload derived from application of the task-independent predictive model trained on data from the mental workload paradigm to text reading data	146
4.5	Results of subject-wise one-tailed permutation tests performed on predictive values	147
4.6	Cumulative predictive values derived from the predictive models of the mental workload classifiers	148
4.7	Predictive value averages of the task-independent predictive model applied to EEG data of each word of the twelve texts, RTLX ratings and incorrectly answered literal text comprehension questions	150
5.1	Experimental timeline of the study presenting pictures from the SMID database	171
5.2	Experimental setup for the SMID paradigm in the lab	171
5.3	Grand average ERPs of morally objectionable, neutral and affectively negative trials from electrode sites Fz, Cz and Pz	174
5.4	Grand average difference curves of the ERPs (1) morally bad minus morally neutral, 2) affectively bad minus morally neutral and 3) morally bad minus affectively bad) at electrode sites Fz, Cz and Pz with permutation test results	175
1	Grand average ERPs of morally objectionable and affectively negative trials from electrode sites Fz, Cz and Pz	225
2	Grand average ERPs of morally objectionable and morally neutral trials from electrode sites Fz, Cz and Pz	226
3	Grand average ERPs of affectively bad and morally neutral trials from electrode sites Fz, Cz and Pz	227
4	Example pictures for morally bad pictures	228
5	Example pictures for morally neutral but affectively bad pictures . . .	228
6	Example pictures for morally and affectively neutral pictures	228

List of Tables

2.1	Overview of frequency bands in the EEG and their corresponding mental states	14
3.1	Classifier calibration error rates of subject-specific classifiers from recording session 1 (includes eye activity)	62
3.2	Classifier prediction errors rates from application of the subject-specific classifiers to testdata from recording session 1 (includes eye activity)	67
3.3	Classifier calibration error rates of the subject-specific classifiers trained on EOG channels from recording session 1	69
3.4	Classifier prediction errors rates from application of the subject-specific classifiers trained on EOG channels to testdata	69
3.5	Classifier calibration error rates of the subject-specific classifiers from recording session 1 (cleaned from eye activity)	74
3.6	Classifier prediction errors rates from application of subject-specific classifiers to testdata from recording session 1 (cleaned from eye activity)	79
3.7	Classifier prediction error rates from application of the subject-specific classifiers from the first recording session to testdata from recording session 2	82
3.8	Classifier prediction error rates from application of the group classifier from the first recording session to subject-specific testdata from recording session 2	85
3.10	Classifier prediction error rates from application of the subject-specific classifiers from the first recording session to all other subjects' testdata from recording session 1	92
3.9	Classifier calibration error rates of the leave-one-subject-out group classifiers from recording session 1	93
3.11	Classifier prediction error rates of applying the leave-one-subject-out group classifiers from recording session 1 to subject-specific testdata of recording session 1	94
3.12	Subject-specific classifier calibration error rates of training with a 50, 32 and a 24 electrode setup on all data from recording sessions 1 and 2	101
3.13	Group classifier calibration error rates of training with a 50, 32 and a 24 electrode setup on all data from recording sessions 1 and 2	102
3.14	Classifier calibration error rates of the subject-specific classifiers from recording session 3 (cleaned from eye activity)	109
3.15	Classifier prediction error rates of applying the subject-specific classifiers from recording session 3 and the group classifier, both trained on 24 electrodes, to subject-specific testdata of recording session 3	116

3.16	Group classifier error rates from the online application scenario of recording session 3	117
4.1	Classification results of the mental workload paradigm	144
5.1	Means and SDs for stimuli categories on the moral acceptance, valence and arousal ratings scales	170
5.2	BCI classification results of the SMID picture based moral evaluation paradigm for all four subjects	177
5.3	BCI classification results of data from Berkum et al. (2009) for all 43 subjects	178
5.4	BCI classification results of data from Leuthold et al. (2014) for all 24 subjects	179

Appendices

- Additional grand average ERP graphs of the three experimental conditions in a paired comparison from the third study regarding monitoring of neural correlates of human moral assessments for future human-machine value alignment, see chapter 5:

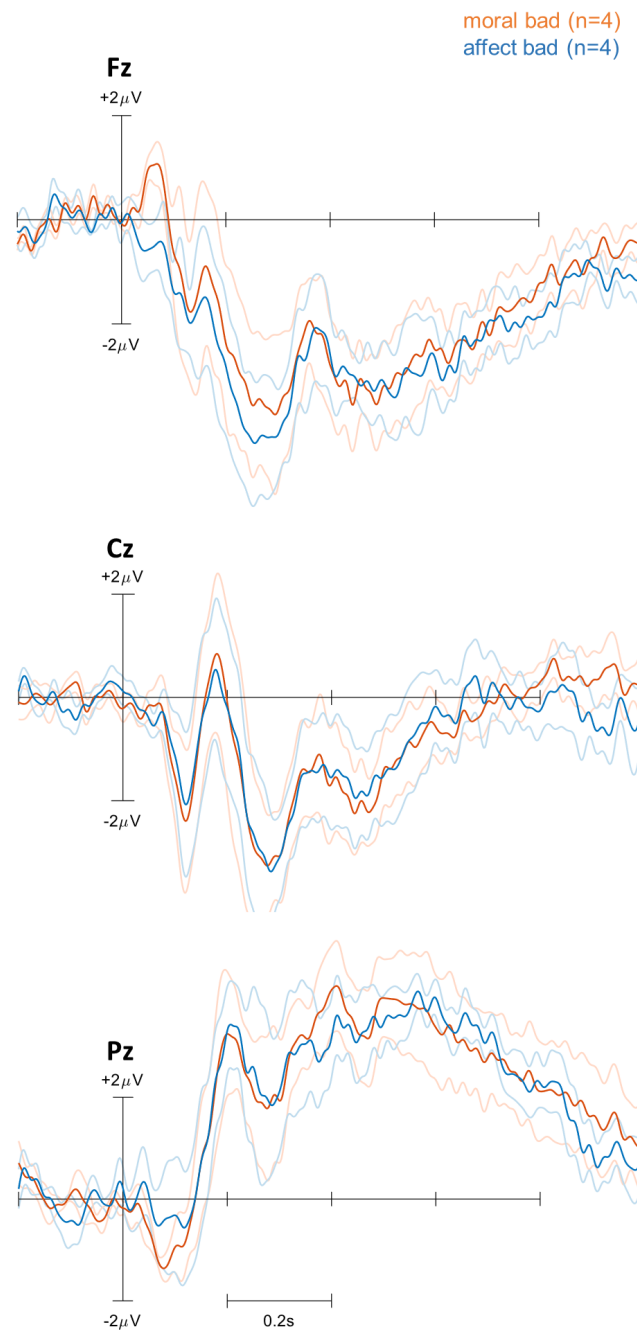


Fig. 1 Grand average ERPs (n=4) of morally objectionable and affectively negative trials from electrode sites Fz, Cz and Pz from the conducted experiment. Standard errors of the mean for the two average curves are plotted in transparent lines.

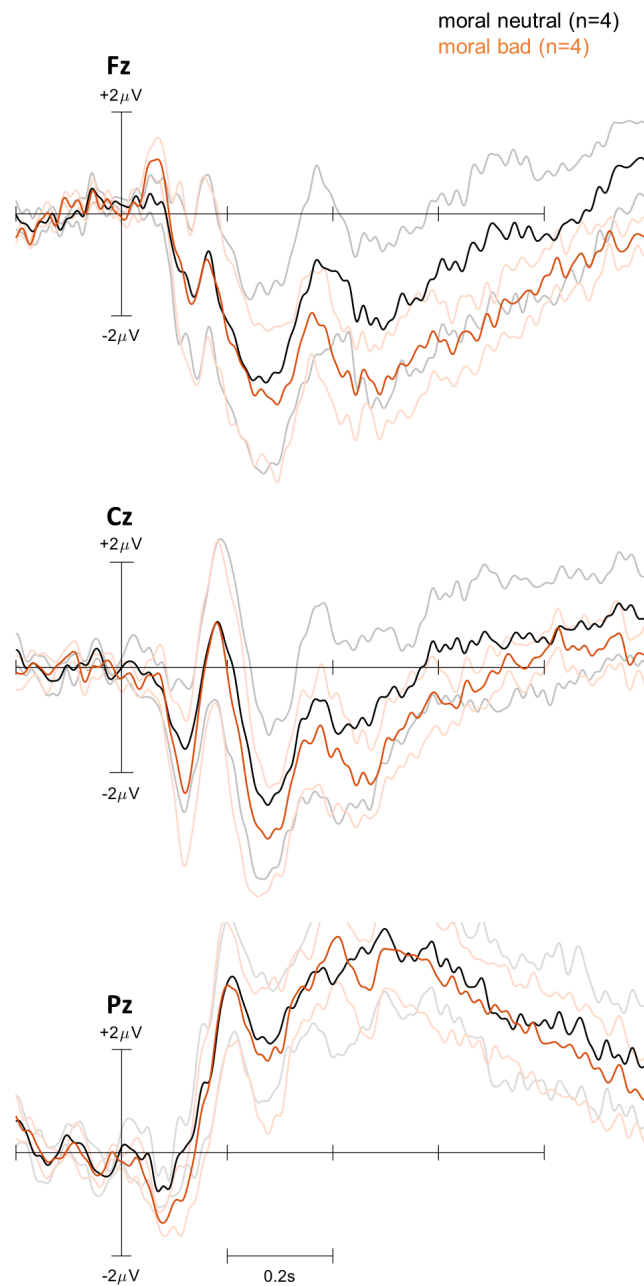


Fig. 2 Grand average ERPs (n=4) of morally objectionable and morally neutral trials from electrode sites Fz, Cz and Pz from the conducted experiment. Standard errors of the mean for the two average curves are plotted in transparent lines.

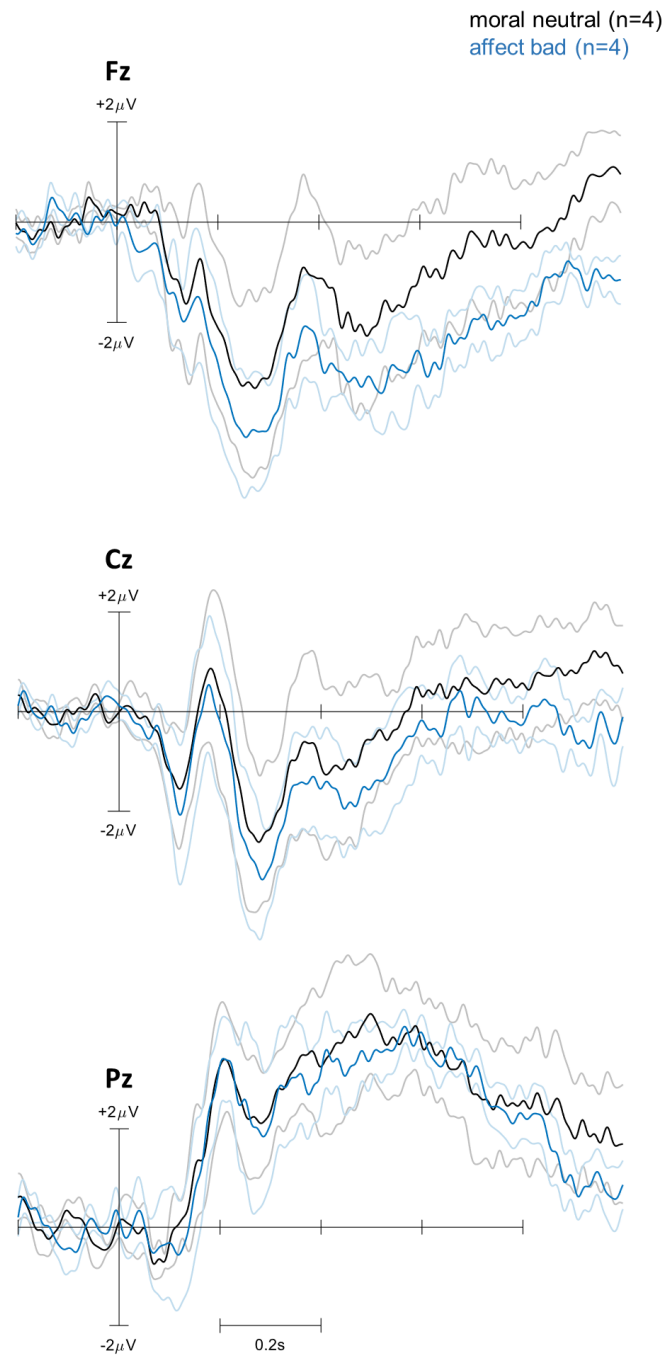


Fig. 3 Grand average ERPs (n=4) of affectively bad and morally neutral trials from electrode sites Fz, Cz and Pz from the conducted experiment. Standard errors of the mean for the two average curves are plotted in transparent lines.

- Example pictures for each picture category generated from the Socio-Moral Image Database (SMID, Crone et al., 2018) used in study three (see chapter 5). Each category consisted of 100 pictures which were shown to participants in randomized order during the experiment.



Fig. 4 Example pictures for the category of morally bad pictures, based on SMID ratings.

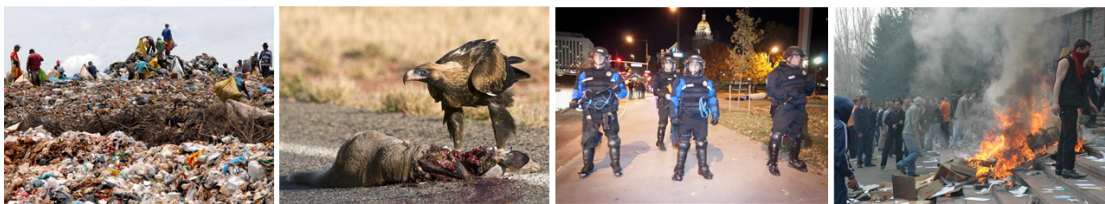


Fig. 5 Example pictures for the category of morally neutral but affectively bad pictures, based on SMID ratings.



Fig. 6 Example pictures for the category of morally and affectively neutral pictures, based on SMID ratings.