Readings in AI 2021
Publications of the ZHAW Centre for Artificial Intelligence
Preface

The ZHAW Centre for Artificial Intelligence (CAI) is a hub for excellence in AI research and application. Our mission is to advance human-centric and trustworthy AI research in Switzerland, thereby providing students with career opportunities in the AI sector, attracting young talent and addressing the great challenges of our time through innovative use of AI.

We believe in the power of interdisciplinary collaboration and engaging in dialogue with the research community, with our students, and with our industry partners. We offer expertise in the areas of Autonomous Learning Systems (reinforcement learning, multi-agent systems, and embodied AI), Computer Vision, Perception and Cognition (pattern recognition, machine perception, and neuromorphic engineering), Natural Language Processing (dialogue systems, text analytics, and spoken language technologies) and Explainable AI (trustworthy machine learning, robust deep learning, and MLOps). With this, the CAI’s vision is to contribute to a society that is worth living in, increasingly supported by AI-driven tools of increased generality that place humans at the centre.

The CAI was founded in April 2021, based on the pre-existing groups of Professors Cieliebak and Stadelmann at ZHAW’s Institute of Applied Information Technology. While the research in machine learning for natural language processing, computer vision and pattern recognition continues, the underlying structure changed, not unlike to repotting a plant, into a new environment. Some of our milestones are depicted in the following timeline:

One structurally new thing is that herewith and for the first time, we issue a research report to give an account of our work. Intended to grow in future editions, this year will focus on the main public results – our scientific publications. Organized by research group, you will find a brief overview of the group’s development over the year 2021, followed by the full text of the published papers of our staff, in the order of their appearance. I wish you, dear reader, an insightful reading.

Winterthur, Spring 2022

Thilo Stadelmann
Director of Centre for Artificial Intelligence
# Table of Contents

## Preface

Table of Contents

## 1 Research Output of the Computer Vision, Perception and Cognition Group  

User adaptation to closed-loop decoding of motor imagery termination ........................................ 5  
The CLAIRE COVID-19 initiative: approach, experiences and recommendations ........... 13  
EEG-Based Online Regulation of Difficulty in Simulated Flying ........................................... 21  
Standardization of neurotechnology for brain-machine interfacing: state of the art and recommendations ................................................................. 33  
Closed-loop EEG study on visual recognition during driving ................................................. 37  
Mobile brain/body imaging of landmark-based navigation with high-density EEG ........... 53  
Invariability of EEG error-related potentials during continuous feedback protocols elicited by erroneous actions at predicted or unpredicted states .............................. 81  
The AI Atlas: Didactics for Teaching AI and Machine Learning On-Site, Online, and hybrid ........................................................................................................................... 99  
A Survey of Un-, Weakly-, and Semi-Supervised Learning Methods for Noisy, Missing and Partial Labels in Industrial Vision Applications ............................................. 123  
XGBoost Trained on Synthetic Data to Extract Material Parameters of Organic Semiconductors ................................................................. 129  
Context-aware learning for generative models ........................................................................ 135  
PrepNet: A Convolutional Auto-Encoder to Homogenize CT Scans for Cross-Dataset Medical Image Analysis ................................................................. 173  
Bias, awareness and ignorance in deep-learning-based face recognition ...................... 181  
Customizing skills for assistive robotic manipulators, an inverse reinforcement learning approach with error-related potentials ......................................................... 195  

## 2 Research Output of the Natural Language Processing Group  

Towards Understanding Lifelong Learning for Dialogue Systems ........................................... 213  
The Sentence End and Punctuation Prediction in NLG Text (SEPP-NLG)  
Shared Task 2021 ........................................................................................................... 221  
ZHAW-CAI: Ensemble Method for Swiss German Speech to Standard German Text ........................................................................................................... 229
Are We Summarizing the Right Way? A Survey of Dialogue Summarization Data Sets ........................................................................................................................................................................................................ 237

Excerpt of conference proceedings: Proceedings of the 6th Swiss Text Analytics Conference (SwissText) ........................................................................................................................................................................................................ 249


Appendix 265

A.1 List of Publications Computer Vision, Perception and Cognition Group ............. 265

List of Publications Natural Language Processing Group .................................................. 267

A.2 CAI Team ............................................................................................................................................... 268

A.3 Location ............................................................................................................................................... 269
1 Research Output of the Computer Vision, Perception and Cognition Group
1 Research Output of the Computer Vision, Perception and Cognition Group

The CVPC group, led by Prof. Stadelmann, conducts pattern recognition research by working on a wide variety of tasks relating to image, audio, and signal data per se. It focuses on deep neural network and reinforcement learning methodology, inspired by biological learning.

Each studied task has its own learning target (e.g., detection, classification, clustering, segmentation, novelty detection, control) and corresponding use case (e.g., predictive maintenance, speaker recognition for multimedia indexing, document analysis, optical music recognition, computer vision for industrial quality control, automated machine learning, deep reinforcement learning for automated game play or building control), which in turn sheds light on different aspects of the learning process. This experience is used to create increasingly general AI systems built on neural architectures.

In 2021, the group consisted of a core of 7 researchers (one full professor, 3 senior scientists, 2 doctoral students and 2 research assistants that pursue a master's degree in part time) and was joined by a research intern, Adhiraj Ghosh, in June. Much of the group's activities revolved around finishing project “FWA: Visual Food Waste Analysis for Sustainable Kitchens” and working towards the expected successful closure of projects “RealScore – Scanning of Real-World Sheet Music for a Digital Music Stand” and “DIR3CT: Deep Image Reconstruction through X-Ray Projection-based 3D Learning of Computed Tomography Volumes” in 2022. Publications from these projects are expected in 2022.

Hence, the group’s publications in 2021 revolve around a diverse set of topics, amongst them highly interesting side projects: Amirian et al.’s work on generative modeling for medical image data homogenization involved the work of two Bachelor students, Jonathan Gruss and Yves Stebler; Dr. Chavarriaga contributed to a publication out of his involvement in the CLAIRE Covid-19 task force; and the paper Prof. Stadelmann co-authored on coupling machine learning and simulation to extract modeling parameters for photovoltaic cells won the best-paper award at the Swiss Data Science Conference 2021; finally, Dr. Chavarriaga’s prior work on neurotechnologies, e.g. EEG signal analysis with machine learning, was published in different prestigious journals.

Besides side projects 2021 saw the successful completion of several studies that had been started years before. The work conducted within project “QualitAI - Quality control of industrial products via deep learning on images” (2017-2020) now led to a survey on methods for dealing with degrading label quality and/or quantity; the work on face recognition in project “Libra: A One-Tool Solution for MLD4 Compliance” (2016-2019) finished with a larger study on the issue of algorithmic bias in face recognition systems. Finally, a multi-year study on didactic methods and their result in terms of student success resulted in a publication of Prof. Stadelmann’s educational concept for on-site, online and hybrid teaching of AI and machine learning that won the 2019 ZHAW best teaching award.

We thank our project partners, students, and funding organizations for their generous support and effort, without which these results (and the results forthcoming in future years) would not have been possible!

The CVPC 2021 team

Thilo Stadelmann, Frank-Peter Schilling, Ricardo Chavarriaga, Javier Montoya, Lukas Tuggener, Mohammadreza Amirian, Pascal Sager, Raphael Emberger, and Adhiraj Ghosh
User adaptation to closed-loop decoding of motor imagery termination

Bastien Orset, Kyuhwa Lee, Ricardo Chavarriaga, Senior Member, IEEE and José del R. Millán, Fellow, IEEE

Abstract — One of the most popular methods in non-invasive brain machine interfaces (BMI) relies on the decoding of sensorimotor rhythms associated to sustained motor imagery. Although motor imagery has been intensively studied, its termination is mostly neglected. Objective: Here, we provide insights in the decoding of motor imagery termination and investigate the use of such decoder in closed-loop BMI. Methods: Participants (N = 9) were asked to perform kinesthetic motor imagery of both hands simultaneously cued with a clock indicating the initiation and termination of the action. Using electroencephalogram (EEG) signals, we built a decoder to detect the transition between event-related desynchronization and event-related synchronization. Features for this decoder were correlates of motor termination in the upper μ and β bands. Results: The decoder reached an accuracy of 76.2% (N = 9), revealing the high robustness of our approach. More importantly, this paper shows that the decoding of motor termination has an intrinsic latency mainly due to the delayed appearance of its correlates. Because the latency was consistent and thus predictable, users were able to compensate it after training. Conclusion: Using our decoding system, BMI users were able to adapt their behavior and modulate their sensorimotor rhythm to stop the device (clock) accurately on time. Significance: These results show the importance of closed-loop evaluations of BMI decoders and open new possibilities for BMI control using decoding of movement termination.

Index Terms— Brain–machine interface (BMI), EEG, motor imagery, motor termination, latency, decoding adaptation.

I. INTRODUCTION

Brain machine interfaces (BMIs) aim at providing communication and control pathways for people with motor disabilities [1]. A BMI bypasses the natural motor pathways, enabling users to control a large variety of external devices and interact directly with their environment, as the BMI decodes users’ intentions directly from the analysis of brain signals and translates intentions into commands for an external device such as spellers [2], [3], avatars [4], robots and wheelchairs [5], [6], hand neuroprostheses [7], [8] as well as for neurogaming and consciousness assessment [9], [10].

This paper was submitted on the 22nd February 2019. A complementary analysis of the data recorded during the offline session reported here appeared in a conference paper [11]. This work was supported by the Swiss NCCR Robotics. B. O., K. L. and R. C. are with the Defitech Chair in Brain-Machine Interface, Center for Neuroprosthetics, Ecole Polytechnique Fédérale de Lausanne, Geneva, Switzerland (email: bastien.orset@epfl.ch). J. d. R. M. is with the Dept. of Electrical and Computer Engineering & the Dept. of Neurology, University of Texas at Austin, USA; he is also with the Defitech Chair in Brain-Machine Interface, Ecole Polytechnique Fédérale de Lausanne, Geneva, Switzerland (email: jose.millan@Austin.utexas.edu).

Most non-invasive BMIs based on voluntary modulations of brain rhythms aim at detecting the initiation of an imagined movement. Hence decoders are usually trained on samples from the periods before and after onset. Once the onset is detected, predefined commands can be triggered. Although the detection of imagined movement initiation is critical in the process, decoding the volitional interruption of motor imagery (MI) is of equal importance in order to endow brain-actuated devices with more natural behavior. While decoding of movement initiation is the focus of multiple works [7], [8], [12], [13], decoding of movement termination has been rarely investigated [14], [15]. Indeed, only two studies have explored so far the feasibility of decoding termination of MI. The first one showed the possibility to build a brain switch using one Laplacian channel [16]. Similarly, Bai et al. [17] investigated a β rhythm-based BMI in repetitive motor imagery.

In this study, we investigate the use of a specific decoder for hand MI termination in a closed-loop BMI. We show for the first time that, during closed-loop operation, BMI users can adapt to their own decoder and compensate for its latency to stop precisely on time. To this end, we designed a task enabling us to capture the correlates of sustained MI as well as the neural correlates of MI termination. This task was inspired by Libet’s experiment on motor initiation [18] and explored in the context of BCI [19].

It is known that changes in the brain rhythms during planning and execution of movements, as well as in the case of MI, can be observed in and decoded from human EEG [20]. After movement termination, an increase of power (event-related synchronization, ERS) is induced in the β band (13-25 Hz). Such synchronization, often called β rebound, can last for about a second. Although the role of β rebound is still under debate, it is currently thought to have a function of inhibition of the motor cortex by somatosensory processing [21], [22]. Oscillatory activity in the β band has been also linked to an active process aiming to maintain the current sensorimotor or cognitive state (i.e., status quo) [23]. Similarly, it has also been reported the presence of an ERS in the μ band (8-13 Hz) that could be interpreted as an electrophysiological correlate of cortical idling state in sensorimotor areas [24]-[26]. Such synchronizations can be explained by an increase of rhythmic activity paradoxically due to a decrease of excitability of cortical neurons or inhibited cortical neurons [27], [28]. This was also reported after MI tasks [28]. When performing hand-related motor tasks, ERS can mainly be observed in the contralateral hand representation area. In the β band, this synchronization can also be seen in the supplementary motor
area (SMA) located in mid-central areas of the brain with slightly higher frequencies and an earlier beginning compared to the contralateral ERS [29], [30].

II. METHODS

A. Experiment setup

Nine healthy naïve subjects (19-26 years, 2 females) participated in the experiment. The study was approved by the Cantonal Committee of Vaud, Switzerland for ethics in human research (CER-VD) and subjects gave their written permission and signed a consent form.

B. Offline Protocol

Participants were comfortably seated in front of a PC monitor and asked to perform kinesthetic MI (i.e., imagining the feeling and asked to perform kinesthetic MI (i.e., imagining the feeling associated with performing a movement) of both hands simultaneously. The total duration of a trial was 13 s. Subjects first fixated the cross in the middle of the screen (3 s), then they performed MI (2 to 4 s), and finally they rested until the hand clock finishes its turn (6 to 8 s). The total time of a clock hand revolution (MI plus rest) was 10 s. During the rest period after MI termination (MI), subjects were instructed to stay calm avoiding any muscular contraction or blinks while the hand finished to revolve around the clock. In between trials, a relaxation phase of 7 s was introduced to allow participants to blink and rest. Fig. 1 illustrates the structure of the trial. Each subject performed 4 runs of 30 trials each (120 trials in total).

![Fig. 1. Trial structure during offline protocol. During a trial, the subject is asked to continuously look at a fixation cross in the center of the clock. The subject is instructed to stay calm for the first 3 seconds without moving or blinking. A clock hand (green bar) indicates to the subject to initiate his motor imagery of both hands. When the clock hand reaches a target (red bar), the subject stops motor imagery and stays at rest (no blink or movement). A period of 7 seconds following each trial allows the subject to relax.](image)

C. Online Protocol

After training a decoder on the offline data (c.f. Section II-E), participants’ task was to stop the clock hand on a target in real-time by terminating their MI action. Participants were asked to adapt to their individual decoder in order to stop precisely on the target. To do so, participants had to evaluate the latency of the BMI output during an initial calibration phase that preceded the actual experiment. The calibration phase consisted of 20 online trials and, afterwards, each subject performed 4 runs of 25 trials each (100 trials in total).

In this protocol, a gauge (not shown in Fig.1) was additionally integrated on the clock hand showing the BMI output as a source of continuous feedback. The clock hand stopped when the gauge was filled ($P_{\text{gauge}} > 1$). The BMI output corresponded to the integration of the output probabilities of the MI decoder to each single EEG sample based on an exponential moving average (Eq.1)[31]:

$$P_t = \alpha P_{t-1} + (1 - \alpha) P_t \quad (\text{Eq. 1})$$

Where $P_t$ is the smoothed likelihood, $p_t$ is the likelihood of detecting MI termination (output of the MI decoder) and $\alpha$ is a smoothing parameter, $0 \leq \alpha \leq 1$.

The smoothing parameter $\alpha$ was individually set for each subject by the operator based on the data of the calibration phase and kept fixed for the entire online evaluation. Table 1 reports the value of this parameter for each subject.

<table>
<thead>
<tr>
<th>Subject</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0.85</td>
<td>0.85</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

To decrease the delay due to the exponential moving average, the smoothed probabilities were passed through a burst framework which was gradually increasing the bar shown to the user. This burst framework was defined by the following equation (Eq. 2).

$$P_{\text{gauge},t} = P_{\text{gauge},t-1} + (P_t - 0.5) \quad (\text{Eq. 2})$$

$$P_{\text{gauge}} \leq 1$$

$$P_{\text{gauge}} - \alpha = 0.1$$

D. Recording System

EEG signals were recorded at a sampling frequency of 512 Hz with 16 active surface electrodes placed over the sensorimotor cortex i.e., on positions Fz, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, and CP4 according to the international 10/10 system (reference: left earlobe; ground: AFz; g.tec gUSBamp, Guger Technologies OG, Graz, Austria). The amplifier was set with a hardware band-pass frequency between 0.01 and 100 Hz (Butterworth 4th order) and a notch filter between 48 Hz and 52 Hz. A common average reference was used on the EEG raw data to enhance the signal-to-noise ratio.

E. Offline Classification

Power spectral densities (PSD) were computed in a 1s-window using the Welch’s method (0.5 s window with 0.25 s non-overlapping window) from 4 to 40 Hz with a 2 Hz resolution on the 16 channels, yielding a total of 304 features. We compared three different classifiers based on different bands: $\mu$ band (8-13 Hz), $\beta$ band (20-26 Hz) and all bands together (4-40 Hz). Then for each of these classifiers, 10-fold trial-based cross-validation was performed where the 6 best features from each fold were selected based on their Fisher Score. Using these features, a Diagonal Linear Discriminant Analysis (DLDA) was trained. Feature vectors were extracted from each sample in the training dataset and z-score normalized. Their mean and variance were applied on the feature vectors in the testing dataset. The movement termination decoders were trained to distinguish between sustained MI and MIt. 1-second long samples were computed with a sliding window (shifted every 62.5 ms) in the time intervals [-2, 0] s (MI) and [0.5, 2.5] s (MIt) with respect to the offset ($t=0$). Using the 1s-long overlapping sliding windows every 62.5 ms within the 2s interval yield 17 samples per trial.
To assess the classification performance, we calculated the accuracy at the sample level over the 10-fold cross validation. The accuracy was defined as the number of correctly classified samples over the total number of samples and was computed for each fold. We estimated the significance threshold above chance level at the 95% confidence interval based on the inverse binomial cumulative distribution (uniform priors, n = 408 samples in test set), leading to a value of 54.17%.

F. Asynchronous classification in online session

Based on the results of the cross-validation, and using all the data from the offline session, we trained a decoder using all the 304 features available (4-40 Hz). After normalization, a total of 6 features were selected based on their Fisher Score ranking. Then, these features were used to train a DLDA classifier to detect MI in real-time during the online session.

G. Pseudo-online analysis

A pseudo-online (PO) analysis was performed on the offline and online data to further study the classifier behavior in real-time at the trial level. To do so, we compare the behavior of the online classifier (online PO) and the behavior of the different classifiers used for cross-validation (offline PO). In this analysis, the classifier was tested in the time interval [-3, 4] s with respect to the offset cue. During this time interval, the likelihood (i.e., the probability of detecting MI) was calculated from the decoder on samples computed with a 1s-window shifted every 62.5 ms. The posterior probabilities were then smoothed as explained in Section II.C. Using this decoder, we measured the decoding latency, which we defined as the time when the average posterior probabilities over trials were crossing the significance threshold above chance level of 54.17%.

III. RESULTS

A. Time frequency analysis

A spectral analysis was first performed on central channels (C3, Cz, C4) where the event-related spectral changes were evaluated for MI onset and offset using Fast Fourier Transform on 1s-Hanning window shifted every 62.5 ms [32]. A baseline period [-2, -1] s with respect to the onset was used to compute the spectral changes. Fig. 2A shows the grand averages across subjects recorded during the offline session. During sustained MI, a desynchronization (ERD) could be observed in the μ and β bands, more prominently on centro-lateral channels (C3, C4) as expected in the case of bilateral hand MI. On the opposite, an ERS was seen on these channels after stopping MI, mainly noticed in the upper μ (11-13 Hz). An ERS was also observed in the high β band (20-30 Hz) mainly in C3 and C4, but it exhibited larger inter-subject variability. Importantly, one can observe that β ERS had a shorter duration than μ ERS (Fig. 2B).

Indeed, β ERS started at about 0.8 s after the offset and lasted for 2.1 s, while μ ERS started later at 1.5 s but remained until the end of the trial (α = 0.05, repeated measure t-tests based on t-statistics, FDR corrected for multiple comparisons).

B. Offline classification

We evaluated the performance of our classifier for decoding MI termination with three different bands: μ (8-13 Hz), β (20-30 Hz) and all bands (4-40 Hz). Fig. 3A shows the accuracy of the three decoders for each subject and in grand average. The μ and β bands-based decoders yielded accuracies of 73.3±7.4% and 73.0±7.0% (mean ± std), respectively; while using all bands reached an accuracy of 76.2±6.4%. This performance improvement using the classifier based on all the bands was, however, not statistically significant on average (one-way ANOVA F(2,24) = 0.57; p = 0.574). Importantly, all three classifiers reached high accuracy and were statistically above the significance threshold (54.17%) for every subject. Additionally, Fig. 3B shows the Fisher scores of the spectral features averaged over the different folds of the cross validation and subjects in the case of the third classifier. Note that most of the discriminant features can be found in channels located over the hand motor area (C4, C3 and CP4) in the upper μ band and, to a lesser extent, in the β band. Later, we used this classifier for the online session.

C. Asynchronous classification

For online session, we trained, for each subject, a decoder such as described in Section II.F. Fig. 4 reports the features that
were selected for the online decoders. Most of the features were located in the upper µ band.

Fig. 5A shows the latency for each subject during the online session when using our integrative framework for increasing the reliability of BMI outputs (see Section II.C). The latency was computed for every trial by calculating the difference between the time when the participant is supposed to stop and when the clock hand actually stopped moving. On average, we obtained a latency of $-0.1 \pm 1.7$ s (median ± diff. percentile).

Fig. 5B reports the number of trials where MI termination was detected—on average, in 85.44% of the trials. For the remaining trials, the clock hand continued rotating for 10 s until the end of the trial. Interestingly, Fig. 5B illustrates that subjects with a high accuracy ($\geq 80\%$) also exhibited a small median latency. These results show that participants were able to control the offset of their MI precisely. More importantly, these results indicate that the latency is consistent over the trials because of a relatively small inter-trial variability in most of the

---

**Fig. 3.** Movement termination decoder. A. Classifier accuracies are reported for every subject (s1-s9) as well as in grand average (GA). The standard deviation is shown on each bar. Dots indicate performance for each subject. The horizontal dashed line corresponds to the significance threshold above chance level (54.17%). B. Fisher score map averaged over all subjects. Fisher scores are shown for the features (channels x frequencies). Higher values (i.e., yellow color) indicate highly informative features while blue colors indicate less discriminant features. The scores were normalized for each subject using min-max scaling.

**Fig. 4.** Feature occurrences for online decoding for all subjects. The number of times a feature was selected for online classifier was counted across subjects. Higher values (i.e., yellow color) indicate highly selected features while blue colors indicate less selected features.

**Fig. 5.** A. Distribution of latency for every trial and for each subject. Each point corresponds to the latency of one trial and was calculated by computing the difference between the time when the participant is supposed to stop and when the clock hand stopped to move. Boxplots illustrate the distribution and the median latency for every subject. B. The graph shows the median latency in function of the percentage of trials where a stop was detected for each subject. (except subject s8).
D. Pseudo-online classification

We further investigated the behavior of our approach, comparing its dynamics between offline and online sessions. To do so, we performed a pseudo-online analysis of the offline session (offline PO), the calibration phase, and each run of the online session (online PO). The results are reported in Fig. 6. On average across all subjects, one can observe that the online PO is shifted back in time, crossing the 54.17% significance threshold above chance level at $t=0 \pm 0.11$ s (mean ± SEM) before offset cue; while the offline PO crosses it at $t=1.25 \pm 0.15$ s (mean ± SEM) after the offset cue as expected. The calibration phase exhibits a similar behavior to the offline PO. Additionally, we can see that the latency was consistent over online runs. These results confirm that subjects learned rapidly to adapt to the latency of their decoder during closed-loop BMI usage and were able to stop the clock hand precisely on time.

E. ERS modulation during online session

We computed the spectrogram in central channels during the online session (Fig. 7). Compared to the offline session (Fig. 2A), one can observe a temporal shift of both ERS in the upper $\mu$ and $\beta$ bands that appear now aligned to the offset cue (Fig. 7A). Interestingly, we can also notice stronger ERS patterns, especially in the $\beta$ band. This may be explained by subjects who have more MI practice. Similarly, we evaluated the timing of $\mu$ and $\beta$ ERS ($\alpha = 0.05$, repeated measure $t$-tests based on $t$-statistics FDR corrected for multiple comparisons). As Fig. 7B illustrates, during the online session, the $\beta$ ERS appeared at $-1.1$ s; while, as before, the $\mu$ ERS started to develop slightly later at $-0.8$ s with respect to the offset cue. Interestingly, ERS in both bands seemed to last until the end of the trials as shown by the significant differences w.r.t to the baseline period (top colored bars in Fig 7B). Additionally, we also computed the spectrogram for channels C3 and C4 at the single trial level (Fig. 8) from every subject in order to investigate the inter-trial variability across the different bands that cannot be observed in the grand average spectrogram. One can notice that the upper $\mu$
ERS is well aligned with the offset cue and is consistent over trials. On the other hand, there is a large variability of the $\beta$ rebound over the trials and, more importantly, it is less well aligned with the offset cue.

IV. DISCUSSION

The present study investigated the feasibility of decoding the spontaneous termination of an imagined movement, based on the natural electrophysiological correlates of such a task. More importantly, we also aimed to answer whether BMI users were able to control the offset of their sensorimotor rhythm in real-time to stop a device accurately at a specific position. We evaluated such an ability by analyzing the BMI performance accuracy and latency.

A. Time-frequency analysis

From our time-frequency analysis, we observed neural correlates of motor termination consistent with literature [20], [27], [28]. These correlates are characterized by an increase of power in the upper $\mu$ and $\beta$ bands. Comparing these correlates, it appears that the $\mu$ ERS is more prominent and more reliable over subjects since it was observed in eight of them, while significant $\beta$ ERS was found in five subjects. Moreover, this $\mu$ ERS tends to last longer than $\beta$ ERS, which likely help in detecting MI termination. This difference of duration can be explained by their functional role. Indeed, $\beta$ rebound is identified as a neural correlate of termination [21]–[23], while $\mu$ rebound is reported as a neural correlate of an idling process [24]–[26]. Because of these different roles, $\beta$ rebound can be seen more as a phasic modulation at offset, whereas $\mu$ synchronization is expected to remain as it is the paradigmatic pattern at rest. Thus, our results are consistent with the current hypotheses regarding the functional roles of both $\beta$ and $\mu$ rhythms [20]–[26]. Importantly, these correlates are found mainly over the sensorimotor cortex (C3 and C4) as well as for some subject in central channel Cz. The location of these correlates is consistent also with fMRI literature for movement initiation and termination [15], [33], [34]. Indeed, premotor supplementary area (preSMA) and premotor cortex were identified as brain region involved in the process of stopping a voluntary action. It is also known that the $\beta$ rebound is mainly observed in premotor cortex as well as supplementary motor area [29], [30]. Hence, these results support the putative role of these correlates in the termination of a motor action.

B. Decoding motor imagery termination

Performance of our decoder were similar to those reported by Bai et al. [17] (average accuracy of $\approx$76% and $\approx$75%, respectively). Although our decoder followed the same strategy, the studies mentioned above focused on the $\beta$ band. In contrast, we found that additional information can be extracted from the upper $\mu$ band (see Figs. 3B and 4). Furthermore, the comparison between decoders based on different frequency bands did not show significant differences in term of accuracies. These observations indicate that BMIs should not be based on pre-selected physiological features that may vary from subject to subject, but should be personalized accordingly –e.g., using feature selection on a broader frequency range as done in the present work.

C. Decoding latency and adaptation

By looking at the offline PO analysis, we can observe that our decoder shows an average latency close to 1.4 s. This latency can be explained mostly by the neurophysiology of ERS. Indeed, ERS develops around 1 s after MI offset, which makes an earlier detection implausible. In contrast, the online evaluation (online PO) exhibits decreased average latency with values close to the MI termination cue. We can thus conclude that, based on the BMI feedback provided by the decoder, subjects were able to adapt and compensate the decoding latency, likely by anticipating the right moment to stop MI.

By looking at the spectrograms in the grand average and at the trial level (Figs. 2, 7, and 8), it appears that the neural correlates of MI termination in the upper $\mu$ band are much more consistent between subjects and show a higher inter-trial stability than in the $\beta$ band. This consistency leads to a robust decoding of motor termination and make the responses of the system more predictable. This probably enabled users to stop accurately on target.

On the other hand, $\beta$ modulations seem also to play a critical role during the closed-loop experiments, supporting that the decoders are detecting MI termination and not just the rest state after the end of the task. Although in the offline condition the performance curve of the detector shows a plateau (Fig. 6), indicating that there might be a strong component of rest, this is not the case once subjects go online and learn to anticipate (and eventually modulate) their brain signals. The sharper detection curve for the online sessions (Fig. 6), decaying about 1 s after $t=0$, suggests that the classifier is not simply decoding the rest state, but a fast transitory EEG correlate associated to MI termination. This correlate is likely the $\beta$ rebound observed in Figure 7, which is initially very prominent and then decays rapidly. The decoders, which use $\beta$ features (Fig. 4), should be detecting the $\beta$ rebound.

V. CONCLUSION

We presented an approach to decode the termination of an imagined movement using both the upper $\mu$ and $\beta$ rhythms. To the best of our knowledge, we are showing for the first time that offset MI decoding has an intrinsic latency mainly due to late appearance of neural correlates of motor termination; however, and critically, this latency can be compensated by BMI users. This compensation is only possible because of the reliability of the decoder as well as the consistency of offset correlates, which makes it possible for users to predict their BMI dynamics. This result also highlights the importance of online evaluation of BMI systems and the implications that closed-loop interactions have in the system performance [4], [35]. The natural scenario for detection of MI termination is its combination with detection of MI initiation for natural control of neuroprostheses. Such an approach would, for instance, allow users to better control the degree of grasping by exploiting two different but associated MI processes, namely MI initiation and termination.
REFERENCES


The CLAIRE COVID-19 initiative: approach, experiences and recommendations

Gianluca Bontempi1 · Ricardo Chavarriaga2 · Hans eD Canck3 · Emanuela Girardi4 · Holger Hoos5 · Iarla Kilbane-Dawe6 · Tonio Ball7 · Ann Nowé8 · Jose Sousa9 · Davide Bacciu10 · Marco Aldinucci11 · Manlio eD Domenico12 · Alessandro Saffiotti13 · Marco Maratea14

Published online: 9 February 2021 © Springer Nature B.V. 2021

Abstract

A volunteer effort by Artificial Intelligence (AI) researchers has shown it can deliver significant research outcomes rapidly to help tackle COVID-19. Within two months, CLAIRE’s self-organising volunteers delivered the World’s first comprehensive curated repository of COVID-19-related datasets useful for drug-repurposing, drafted review papers on the role CT/X-ray scan analysis and robotics could play, and progressed research in other areas. Given the pace required and nature of voluntary efforts, the teams faced a number of challenges. These offer insights in how better to prepare for future volunteer scientific efforts and large scale, data-dependent AI collaborations in general. We offer seven recommendations on how to best leverage such efforts and collaborations in the context of managing future crises.

Keywords Artificial intelligence · COVID-19 · Emergency response

Introduction

Inspired by successful early use of AI by China, Taiwan, Singapore and South Korea to support the management of the COVID-19 pandemic, on 20 March 2020 CLAIRE, the Confederation of Laboratories for AI Research in Europe (CLAIRE) launched a volunteer effort to help tackle the pandemic. As the World’s largest, non-profit network of AI researchers, CLAIRE was quickly able to recruit 150 volunteer AI researchers. This report describes the major activities and achievements of these volunteers, and shares experiences, lessons learnt and recommendations.

The starting point for CLAIRE’s COVID-19 initiative was the insight that the AI community has much to offer in support of efforts to handle the pandemic, its societal and economic consequences, and many AI researchers and practitioners stood ready to help public institutions in the front line of the crisis (Luengo-Oroz et al. 2020).
was that AI could be successfully used across a broad spectrum of areas directly related to managing the COVID-19 crisis, such as:

- Analysis of existing drugs to test their efficacy against COVID-19
- Analysis of data from patients in intensive care, to support prioritisation in triage and therapy
- Analysis of epidemiologic and mobility data, with the goal of better modelling and predicting the spread of the virus, and of facilitating the assessment of impact of containment actions
- Use of advanced 3D printing approaches, with the goal of alleviating the scarcity of equipment for protection and intensive therapy
- Use of automated scheduling and resource management approaches, with the goal of efficiently managing scarce resources in the medical sector (ICU beds, ventilators, specialists) and other key elements of public infrastructure (personnel, warehouses).

These and many other examples suggest that AI techniques can play a key role in assisting human experts with managing the pandemic and its economic aftermath. We note that, as evident even from the small set of examples given above, it is clear that a broad spectrum of AI techniques and approaches can be brought to bear; for this reason CLAIRE, whose research network spans all areas of AI, across all of Europe, saw itself as particularly well-positioned to mobilise bottom-up support for the use of AI techniques and expertise in fighting the pandemic and in managing its impact on societies across Europe and the world.

**Setup phase**

Directly after CLAIRE’s COVID-19 initiative was launched in late March 2020, a task force was put into place to coordinate the effort and the volunteer experts supporting it.

This task force collected information on the various initiatives on leveraging AI techniques in the context of COVID-19 and supported the development of new projects, connecting the European network of AI experts together with health institutions and governments. By the end of March, the task force had enrolled 150 volunteers, covering the full spectrum of AI methods, tools and technologies. Volunteers indicated their willingness to work on one or more of 11 research topics. Of these, a significant number of volunteers and topic team leaders were found for 7 topics:

- Epidemiological data analysis—10 volunteers
- Mobility and monitoring data analysis—36 volunteers
- Bioinformatics (protein and molecular data analysis)—25 volunteers
- Image analysis (CT scans, X-ray)—47 volunteers
- Social dynamics and networks monitoring—8 volunteers
- Robotics—5 volunteers
- Scheduling and resource management—30 volunteers

**AI & COVID-19 resource database**

An initial challenge was to ensure volunteers were aware of other initiatives already underway globally. By crowdsourcing volunteers’ efforts, a catalogue of AI & COVID-19 related resources was assembled. Although not designed to be exhaustive, this catalogue now lists 129 resources, covering:

- Funding opportunities (21 calls still open at time of drafting)
- Datasets (30 resources)
- Hackathons, challenges and webinars (7 listed)
- Other initiatives (71 listed)

To enlarge the list of initiatives and provide additional information to our volunteers we linked our database to the following catalogues of funding and initiatives:

- GovLab repository
- COVID-19 research funding monitor
- Coronavirus Funding Opportunities
- NIH Open-Access Data and Computational Resources to Address COVID-19

**Overview of research activities**

The 7 groups of volunteers, led by the topic coordinators and with the support of the task force team, are working on several outcomes summarised below.

**Epidemiological modeling and decision support**


No. of volunteers: 10.

This research group works on different types of models for epidemics (Pernice et al. 2020; Report 9: Impact of non-pharmaceutical interventions (NPIs) Response Team xxxx; Data Science Institute and UHasselt xxxx), ranging from high level compartment models to agent based models, and how they can be used to study the dynamical aspects to improve complex decision taking on the effectiveness of prevention strategies. On the one hand, this involves model fitting and optimisation, on the other hand, learning and optimisation of prevention strategies, using epidemiological...
models as simulation environments (Libin and Guiding, xxxx).

Work is underway to identify collaboration mechanisms and structures, considering the support AI can offer in decision-making. This recognises the multi-criteria nature of the problem, balancing the needs of different stakeholders all of whom should be involved.

**Mobility and monitoring data analysis**

Topic coordinator: Jose Sousa, Faculty of Medicine, Health and Life Sciences, Queen’s University Belfast, Northern Ireland.

No. of volunteers: 36.

This work sets out to understand the symptoms progression through self-reported data and its integration with mobility to forecast healthcare decision making. The goal is the development of an AI multilayer learning approach capable of creating evidence based knowledge, using complex networks for self-supervised learning (LeCun et al. 2015), spatial temporal analysis and deep learning.

Work is underway to understand the data collected under the several self-reporting systems (Sun et al. 2020) and test how useful the self-reported data is to forecast events (Real-time tracking of self-reported symptoms to predict potential COVID-19 2020). The initial models will be produced using different methodologies and compared with the officially reported statistics.

**Bioinformatics (protein and molecular data analysis)**

Topic coordinator: Davide Bacciu, Computational Intelligence and Machine Learning Group, Universita’ di Pisa, Italy.

No. of volunteers: 35.

Work on this topic aims to (1) support the community in characterising the disease from its related structural information, including prediction of viral protein folding; (2) study the interactions between the virus and human hosts, including analysing protein–protein interaction data; (3) design and validate methodologies for filtering, retrieval, and generation of targeted drugs leveraging molecular and well as proteomic information; (4) deliver predictive insights onto the genetic features of the virus.

As a first contribution to the community, the workgroup has created a curated collection of COVID-19-related datasets useful for drug-repurposing tasks, integrating data from multiple studies (Cheng et al. 2019; Ashburner et al. 2000; Janet Piñero et al. 2019; Rose Oughtred et al. 2019) and releasing it as a network comprising protein interactions (Cheng et al. 2019), viral-host interactions (Rose Oughtred et al. 2019), genomic information (Ashburner et al. 2000) and drug interactions (Cheng et al. 2019). This resource has already been released to the community. The group will use the resource to provide a methodology for fast retrieval of drugs whose action can be correlated to target proteins, by leveraging deep learning for graphs (Bacciu et al. 2020).

**Image analysis (CT scans)**

Topic coordinator: Marco Aldinucci, Computer Science Dept, University of Torino, Italy.

No. of volunteers: 48.

Research in this area aims to (1) distil the current state of the art of methodologies and data sets for AI-assisted diagnosis of COVID-19 by way of imaging (TC Scan, X-ray, etc.), with the goal of making diagnosis faster, cheaper and more manageable in the hospital processes (e.g. using low-resolution images); and (2) to contribute to the improvement of multidisciplinary knowledge by cross-breeding knowledge in computer science and radiology aiming at creating better, more informative reference datasets, together with data-gathering strategies, beyond the current outbreak (Tartaglione et al. 2004; Shi et al. 2020).

The team is developing a review paper and contributes to already active projects, including EU H2020 DeepHealth and EU ERDF HPC4AI. Two further projects are motivated by the strongly perceived need to distil science from the hype COVID-19 induced in different aspects of everyday life, including scientific works (Deephealth project: EU ICT-2018; EU ICT-2018).

The first addresses a reproducibility and benchmarking task: the main publically available deep neural networks and datasets will be collected and cross-validated to compare them across a common baseline. This task will need a substantial human and compute effort. For this, the group is finalising an agreement with the Italian National Supercomputing Center CINECA that will actively support the group activity, which will require both training and inference of the cartesian product of networks, datasets and network parameters. A non-trivial but enabling aspect of the work will be designing and experimenting tools making it possible to bring AI workload to supercomputers and make AI experts efficiently use large scale platforms (Aldinucci et al. 2018; Colonnelli et al. 2002).

The second seeks to consolidate AI performance metrics for both datasets and networks, which will be needed to assess both quality and compute efficiency aspects.

**Social dynamics and networks monitoring**

Topic coordinator: Manlio De Domenico, Head of Complex Multilayer Networks Lab—Fondazione Bruno Kessler (FBK), Italy.

No. of volunteers: 8
This work uses AI models to analyse social media data together with social, behavioral and economic data for two main purposes: (1) monitor social dynamics to analyse the COVID-19 “infodemic” – “an over-abundance of information – some accurate and some not – that makes it hard for people to find trustworthy sources and reliable guidance when they need it” (WHO - Novel Coronavirus (2019) – with the goal of identifying, monitoring and analysing the overload of unreliable information; of collaborating with data providers to obtain free access to relevant data; and of creating an interdisciplinary hub of experts to fight the “infodemic”; and (2) develop early-warning signals to support policy, informed by spatio-temporal analysis of emotions and sentiments; quantifying and modelling the socio-behavioural response. Social media are playing a crucial role for spreading information, both reliable and unreliable, during the COVID-19 pandemic. Efforts are devoted to unravel the role played by both humans and software-assisted (i.e., social bots) in disseminating false or inflammatory content for social manipulation, a phenomenon recently discovered during political events (Stella et al. 2018), with the ultimate goal of attracting or driving collective attention (Domenico and Altman 2020) towards a specific information. The group maintains a catalogue of robotic offers and demands relevant to the COVID-19 emergency, and it is liaising active research laboratories across Europe. We have found that the liaison aspect is especially important during a crisis, when access to laboratory resources and material may be seriously limited. It is also supporting euRobotics (the association of European robotic stakeholders) in writing a white paper on the potential usage of robotic technology in the COVID-19 emergency.

**Scheduling and resource management**

Topic coordinator: Marco Maratea, Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi. University Genova, Italy.

No. of volunteers: 30.

The group working on this topic has focused on automated planning and scheduling, and resource management in healthcare systems leveraging AI (deductive) methodologies and tools. An initial assessment of relevant resources has been completed, and a review of relevant publications, data and projects is underway. In addition, collaboration with the Galliera hospital in Genova, Italy, is underway to assist with workforce scheduling and automated planning of the utilisation of operating rooms with scarce resources and equipment (Alviano et al. 2018; Dodaro and Galàtà 2019).

**Recommendations for future efforts in a crisis**

Unfortunately, it is more than likely that our societies will be confronted in the not-so-far future with other crises of similar scale. The results of our efforts thus far demonstrate that rapidly assembled volunteer efforts including large teams of experts, although complex to initiate and coordinate, can make valuable contributions in this context. Important initiatives may result from bottom-up efforts, which may consolidate in white papers, joint project applications and dissemination of annotated datasets. However, preparedness for such future events can be improved in a number of ways, and there are important lessons to be learned. Here, we outline some of our experiences and cautiously formulate some recommendations based on these.

**Involving domain experts and public authorities is difficult in a crisis**

Many domain experts in public and medical authorities were already preoccupied with tackling the pandemic, limiting their scope to assist volunteer teams. As a result, teams had
to develop their own analysis of the problems to tackle, seeking to engage with experts later in the development process as the crisis began to be controlled. The teams who most rapidly developed research outputs were those where the topics being worked on were close to their existing expertise.

Recommendation 1 Effectively interfacing with domain experts and public authorities in a crisis situation is challenging, but this should not discourage qualified volunteers.

Recommendation 2 The contribution that voluntary expert teams can make should be taken into account in planning for future crises.

See also Recommendation 6 below.

The need for open licenses and standards for data

The ethical implications of processing medical and other sensitive data, and the strategic and policy impacts of research during a crisis pose major challenges. While researchers were fully committed to respecting European citizens’ privacy, in accordance with European values, fundamental rights and regulations such as GDPR, the weak standardisation of the data collected on COVID-19 and embryonic state of open data access in the medical and epidemiological field made it difficult to compile bigger data sets needed for the data-driven approaches. Thus ethical, data management and standardisation efforts should be carefully considered from the outset of future volunteering efforts.

Open licenses designed by Creative Commons have been used for several of the products from this effort to encourage reuse. CLAIRE has more broadly analysed the Creative Commons Open COVID Licence, and welcomes this open approach to sharing research products. Accelerating the development of open licenses and standards for medical data and models, such as epidemiological models, and applying them consistently would reduce these challenges in future.

Recommendation 3 Address and coordinate ethical issues, standardisation and data management at the beginning of the research effort. Consider using open licences to support and accelerate data availability.

The need for large open datasets and infrastructures

Many AI techniques, notably from the area of machine learning, depend on access to large-scale data. Work with platforms like Twitter and Facebook, and with mobile telecommunication service providers, should become more routine in the future to help speed up the large-scale analyses required to inform policy based on quantitative measures of human behavioural responses to the pandemic. This also demonstrates the need for a European data space, such as that proposed in the European Data Strategy and the European Open Science Cloud, which should include such datasets. Of course, it is very important to not only ensure the quality of such data sets, but also to protect the rights of citizens, in particular their privacy.

Recommendation 4 Support the development of a European data space and an open data approach to medical and sensitive data for scientific purposes, while protecting individuals’ anonymity, dignity and human rights.

The role of large scale research infrastructures

The ability of large-scale research and development infrastructures, such as the Robert Koch and Francis Crick Institutes (or, indeed, Apple and Google), to redeploy expertise to work effectively on the pandemic is notable. They have offered public authorities single points of contact for key expertise and helped rally efforts of related communities.

Many scientists in AI have shown they are eager to dedicate significant time and effort to voluntary activities which might not be necessarily sustainable in a short-time horizon or according to more conventional funding channels. The CLAIRE initiative purposefully decided to build on this to go beyond a sterile communication exercise, to bootstrap a number of concrete scientific collaborations on a voluntary basis. While this effort has demonstrated that volunteer efforts can be effective, this observation supports the case for a large-scale investment into an AI hub (or lighthouse centre) in Europe, acting as the reference point of European nations and institutions for all AI research and development.

Since the largest part of the research community kept on using conventional means and strategies to deal with the COVID-19 crisis—notably: competitive research funding, publication (though accelerated by a largest use of arXiv distribution services) and networking—a European centre for AI could promote an innovative approach to research collaboration, funding and dissemination.

Fast international uptake was only possible thanks to existing research networks and network organisations, or built upon personal international academic networks or ongoing project consortia. Outreach and links to initiatives across the EU network were difficult to deploy or to set up, given the grass-roots nature of most initiatives.

A European network would establish permanent relations with all the relevant AI institutions and initiatives globally.

Recommendation 5 Establish a European hub (or lighthouse centre) for AI of very substantial scale.
Bridging communication between medical and AI expertise

The software platforms and hardware with suitable computational power needed to make use of advanced AI techniques (as recommended by AI experts) are lacking in many hospital environments, whether for e-Health or other solutions. While the function of hospitals is first-and-foremost to deliver medical care, encouraging the future development collaborations between local hospitals and AI researchers and investment into infrastructure that enables these collaborations during normal times would help reduce these barriers in future crises. The approach can be extended to other areas using, for example, national risk registers that identify topics of concern to build networks with those who have to manage crises.

Recommendation 6 Set up stable collaboration between hospitals and AI researchers, and other areas of work where future crises can be expected.

Organising large teams remotely

Remotely organising teams of as many as 47 volunteers to quickly decide on research priorities and means of delivering that research presents its own challenges. Preventing fragmentation that dilutes effort, documenting research plans and the work underway, communicating effectively within the group and disseminating results become significant overheads that are not easy to resolve using slow, traditional methods. There is a rich selection of tools available to address many of these issues, used especially in the software industry, but familiarity with these tools within the scientific community varies greatly. This pushes teams towards the simplest, lowest-common-denominator, legacy solutions as well as towards pre-existing networks and project teams. While even simple videoconferencing and document sharing tools have enabled substantial work to be progressed and completed, scientific researchers should build their familiarity with complex and feature-rich collaboration tools that now exist. This will improve inter-institutional research, assisting both building of teams, internal collaboration, and dissemination of results.

Recommendation 7 Scientific researchers should become fluent with the collaboration tools and techniques routinely used, for example, in the software industry.

Conclusions

As the COVID-19 pandemic and its wider ramifications have yet to play out, it is far too early to draw any definitive conclusions. But the outcomes of CLAIRE’s COVID-19 initiative suggest that bottom-up, expert-driven, non-profit endeavours can play an important role. It also offers further evidence supporting the basic premise that AI can and should play a key role in handling crises such as this one, both in the health and medical aspects of the COVID-19 pandemic, and in the societal and economic recovery to come. The highly interdisciplinary nature of AI makes it an ideal discipline to create bridges with other scientific domains to attack important societal problems and crises.

In crises resources are limited, time is of the essence and the consequences of action or inaction are severe but difficult to predict. We are convinced that AI, with its potential to support human analysis, planning and decision making has much to offer not just in the context of the current pandemic, but also for handling future crises. We have shown that many experts are willing to work quickly together on novel solutions for the benefits of society. Such collaboration depends on quick access to large amounts of data and information, computation and, most importantly, to each other even when social distancing measures severely restrict physical interaction.

We remain aware that when used, developed and deployed under the pressure of exceptional circumstances, AI technology can be a double-edge sword. It is technologically quite easy to put in place systems that might be difficult to dial back once the crisis is over, eroding privacy and other fundamental rights aided by advanced AI tools and techniques. We must not allow this to happen and must develop standards and frameworks that permit rapid progress without eroding human dignity.

Especially in times of crisis, we need to keep our eyes and resources firmly on AI that enhances human intelligence, helps us recognise and avoid our biases and limitations, and that is designed and used to protect and further our interests as individuals and societies, managing potential risks and reinforcing our European values and the goal of developing human-centred AI.

Despite its limited scale, the experience of the CLAIRE COVID-19 initiative has not only made concrete progress on COVID-19 problems, but offered insights on the potential and limitations of a non-conventional, voluntary and bottom up approach based on the good will of AI experts, fully aware of the societal role of their knowledge and of the importance of an open dissemination of science and contributions to open scientific data collaborations with all stakeholders of our innovation ecosystem.

Additional information on the CLAIRE COVID-19 initiative can be found in the website https://covid19.claire-ai.org/

References


DS - Data Science Institute UHasselt - What kind of models do we use? (2020)


Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demandImperial College-COVID-19 Response Team (2020)


Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
EEG-Based Online Regulation of Difficulty in Simulated Flying

Ping-Keng Jao, Ricardo Chavarriaga, Senior Member, IEEE, and José del R. Millán, Fellow, IEEE

Abstract—Adaptively increasing the difficulty level in learning was shown to be beneficial than increasing the level after some fixed time intervals. To efficiently adapt the level, we aimed at decoding the subjective difficulty level based on EEG signals. We designed a visuomotor learning task that one needed to pilot a simulated drone through a series of waypoints of different sizes, to investigate the effectiveness of the EEG decoder. The EEG decoder was compared with another condition that the subjects decided when to increase the difficulty level. We examined the decoding performance together with behavioral outcomes. The online accuracies were higher than the chance level for 16 out of 26 cases, and the behavioral results, such as task scores, skill curves, and learning patterns, of EEG condition were similar to the condition based on manual regulation of difficulty.

Index Terms—EEG, Real-time decoding of difficulty, Closed-loop adaptation, (Simulated) Flying

1 INTRODUCTION

Engagement in the task is a crucial factor for learning and to facilitate positive outcomes [1], [2], [3]. One way to keep the engagement is keeping the task at a proper difficulty level for the person’s skills. This can be formalized by the theory of challenge point that outlines that the dependency between task performance or potential learning outcome and the task difficulty follows an inverted-U curve [4]. Easy tasks may disinterest people while high levels of difficulty will lead to frustration; both cases hinder the learning process. As a result, the performance or learning outcome is optimized if one can operate at the best difficulty level, a condition referred to as the sweet spot.

A specific aspect in the theory of challenge point is that the independent variable is functional task difficulty. The term functional indicates that the task difficulty should be subject-dependent, because novices and experts behave differently on a task with the same objective task difficulty. Therefore, a continuous adaptation of the difficulty level is essential to keep a person operating around their sweet spot. Many applications, such as education and gaming, may benefit from operating around the sweet spot. However, estimating the sweet spot is not trivial because the subjective difficulty level is time-varying and is a function of current skills, cognitive states, and affective states [4]. This assessment usually requires human intervention, either from the learner herself who should concentrate on learning or from an experienced instructor who is rare. An alternative is to automatically estimate their subjective difficulty.

We hypothesize that physiological signals correlates with the functional task difficulty and utilizing such correlates helps stay in the sweet spot. Therefore, we aim at building a Brain-Machine Interface (BMI) regulator, able to estimate the functional difficulty and, to adapt the difficulty level. We test the feasibility of building such a BMI with a visuomotor task as they are at the core of a wide range of applications. In particular, our task allows subjects to practice their piloting skill with different functional difficulty levels defined by their skills.

During interaction, either the BMI regulators or learners manually decide whether to increase the difficulty level or keep it unchanged. The reason is that, for learning, it is common not to decrease the difficulty level during a session because user’s skills are unlikely to decline. Human training literature also supports our choice; the difficulty level in learning should remain unchanged and increase only when considered necessary [5]. Importantly, adaptively increasing the difficulty level has been shown effective in promoting learning, while increasing it at a fixed rate was not [5]. Therefore, the presented functional difficulty level increases only when the current level is considered to be easy.

In this study, we compare the subject’s performance when the difficulty level was set by their explicit inputs, or automatically regulated by the BMI. In addition, a critical question is to uncover the neural correlates of the subjective difficulty level. We do so by analyzing the features in the offline session. Regarding the evaluation of the BMI against the manual condition, subject performed the piloting task in two sessions, each session with the two conditions. Then, we compare how well learners performed (task score), how their skills evolved (skill curve), and how the difficulty level changed over time (learning patterns). Our results confirm the feasibility of a BMI regulator to support skill learning as its performance is not significantly different than in the
induce expected oscillations of the difficulty levels. Szafir et al. introduced a theoretical framework to analyze the relationship between workload-relevant EEG frequency features, where the task performance to physiological signals [7]. Physiological signals used to decode workload include heart rate [8], [9], pupil dilation [10], [11], [12], electrodermal activity [13], [14], EEG signals [6], [10], [15], [16], [17], and Functional Near-Infrared Spectroscopy (fNIRS) [18]. Among different signals, we particularly focus on EEG because of their better time resolutions and latency compared to other signals.

Although many researchers have been working on decoding workload/difficulty from EEG signals, only a few studies conducted closed-loop experiments. To the best of our knowledge, Pope et al. conducted one of the earliest online experiments using EEG signals and reported workload-relevant EEG frequency features, where the task was Multi-Attribute Task Battery [19]. Their online evaluations mainly concerned whether each index of interest can find the proper timing for an educational robot to attract learners’ perception of difficulty has a strong correlation to workload level [6]. For decoding workload level, there already exist several studies utilizing measurements from task performance to physiological signals [7]. Physiological signals used to decode workload include heart rate [8], [9], pupil dilation [10], [11], [12], electrodermal activity [13], [14], EEG signals [6], [10], [15], [16], [17], and Functional Near-Infrared Spectroscopy (fNIRS) [18]. Among different signals, we particularly focus on EEG because of their better time resolutions and latency compared to other signals.

Although many researchers have been working on decoding workload/difficulty from EEG signals, only a few studies conducted closed-loop experiments. To the best of our knowledge, Pope et al. conducted one of the earliest online experiments using EEG signals and reported workload-relevant EEG frequency features, where the task was Multi-Attribute Task Battery [19]. Their online evaluations mainly concerned whether each index of interest can induce expected oscillations of the difficulty levels. Szafir et al. tried to detect a sudden drop of engagement, in order to find the proper timing for an educational robot to attract the attention of a student [17]. However, the result was not significantly better than a control group. Yusel et al., although not using EEG signals, combined both fNIRS and pupil dilation [10], [11], [12], electrodermal activity [13], [14], EEG signals [6], [10], [15], [16], [17], and Functional Near-Infrared Spectroscopy (fNIRS) [18]. Among different signals, we particularly focus on EEG because of their better time resolutions and latency compared to other signals.

Although many researchers have been working on decoding workload/difficulty from EEG signals, only a few studies conducted closed-loop experiments. To the best of our knowledge, Pope et al. conducted one of the earliest online experiments using EEG signals and reported workload-relevant EEG frequency features, where the task was Multi-Attribute Task Battery [19]. Their online evaluations mainly concerned whether each index of interest can induce expected oscillations of the difficulty levels. Szafir et al. tried to detect a sudden drop of engagement, in order to find the proper timing for an educational robot to attract the attention of a student [17]. However, the result was not significantly better than a control group. Yusel et al., although not using EEG signals, combined both fNIRS and task performance to improve learning experience of piano [18]. The beginners in learning piano had benefitted more in both self-experience and task performance, but this was not necessarily the case for the experienced players. Their approach, however, can hardly be extended to other cases that the system is extremely tailored. For example, each subject and each objective, instead of subjective, difficulty level has an individual threshold that was defined from many pilot studies and also involved brain data from her own training task. Ewing et al. investigated the trends in EEG signals by varying the game demand and mental effort. The trends were then modeled to adapt the level of Tetris [15]. The EEG conditions yielded unstable changes of levels and worse immersive experience than self decisions. A potential way to improve is incorporating advanced decoders because they only compared the $\theta$ band (4-8 Hz) at Fz and upper $\alpha$ band (10.5-13 Hz) at P4 with a baseline activity. Aricò et al. tried to adaptively trigger an assistive system to reduce workload level when overloaded, but the adaptive system was not significantly better than without the system [16]. Walter et al. compared the learning outcomes in octal arithmetics questions between an EEG adaptation system and an error-adaptive system that adjusts the level based on the number of wrong answers [20]. The EEG system was competitive but needed a long observation window of about 80 seconds. Recently, Faller et al. provided feedback of arousal from EEG signals to subjects [10]. The arousal state can be linked to task difficulty, because arousal is the independent variable of the Yerkes-Dodson curve that has the same shape as the theory of challenge point [21]. In the study of Faller et al., subjects had arousal feedback when performing an engaging task that is different from usual learning process, because the task does not tolerate an error [10]. They showed that consciously self-regulating arousal toward lower levels improves task performance. Nevertheless, low-skilled subjects were excluded after screening. Although some studies had positive outcomes, there is still space to improve. For example, although Walter et al. yielded promising results [20], it is still unclear whether the decoder can function with a much shorter observation window and whether using an EEG decoder is still competitive in applications that need a short latency for feedback. Although Faller et al. used a similar task as our study [10], their task was less realistic; their subjects only had one degree of freedom in piloting, and the task stopped after one mistake – contrarily to usual learning practice. In addition, most works only examined the behavioral outcome without evaluating the decoding accuracy during online interaction. Our study design considers all these issues.

3 MATERIALS

3.1 Participants

Thirteen subjects (eight females; Mean age 22.6; SD 1.04) participated in the study. The protocol was approved by the local ethical committee and all the subjects provided written consent. All subjects had normal or corrected-to-normal vision and reported no history of motor or neurological disease, except one subject who previously experienced vasovagal syncope but not during this study.

3.2 Recording Setup

We used a Biosemi ActiveTwo amplifier to record EEG and Electrooculography (EOG) signals at 2,048 Hz. Sixty-four EEG electrodes were placed according to the international 10-10 standard. Three additional channels were placed in the middle of the eyebrows and on the bulge bones below the outer sides of canthi to measure the EOG.

Subjects sat comfortably in front of a twenty-inch screen showing the protocol with 1680x1050 resolution. They hold a joystick (Logitech Extreme 3D Pro) to provide inputs to the protocol. One female subject was left-handed but reported being comfortable using a right-handed joystick.

3.3 Task

The subjects were instructed to pilot the simulated drone through a series of circular waypoints. The subjects controlled the roll and pitch while the drone had a constant velocity, 11.8 arbitrary unit (A.U.) per second, when flying straightly. Figure 1 is a screen-shot of the protocol. The simulated environment was implemented with Unity (https://unity3d.com/). The green circle is the current waypoint to fly through while the next one appears in blue. The purple cross at the center was used to determine whether the drone was inside (hit) or outside (miss) a waypoint when passing by. In either case, the current waypoint disappears and the next waypoint becomes the green one. There were
decoding accuracy and comparing the behavioral outcome sessions. The online sessions aimed at both evaluating the necessary data to build a subject-specific decoder for the online three months. The offline session aimed at collecting necessary. This allows to properly compare the behavioral outcome between EEG decoder-based and Manual (self-paced decisions) interactions.

For all sessions, waypoints of a trajectory were arranged in a way that the subject needed to either perform a pitch or roll every other two waypoints; the waypoints in between were placed 32 A.U. (at least 2.7 seconds) in front of the previous one for letting the subject to adjust the orientation of the drone. The numbers of required pitch and roll maneuvers were balanced for all the directions. The Euclidean distance between waypoints was 32 and 24 A.U. away for pitch and roll, and at least needed 2.7 and 2.0 seconds, respectively.

3.5.1 Offline Session

During the setup, each subject familiarized with the protocol by at least 10 minutes of piloting with a pre-defined training trajectory of 121 waypoints (c.f. Section S.I), each piloting lasted about five minutes. The result of last piloting would be used to personalize difficulty levels (as presented in Section 3.4), and the subjects were informed.

Before the piloting task, we recorded one-minute of EEG signals with eye closed and opened as an EEG baseline for calibration (to be addressed in Section 4.1). Then, each subject piloted through thirty-two trajectories, each having a constant difficulty level. As depicted in Figure 2, the level decreased from level 16 to level 1, and then increased from 1 to 16. Each trajectory was composed of thirty-two waypoints and lasted around 90 seconds.

After each trajectory, the subjects reported a numeric level between 0 and 100 for the assessment of perceived difficulty level. The mean values and standard deviations across subjects were depicted in the right side of Figure 2. They also declared whether the trajectory was easy (green dots), hard (blue dots), or extremely hard (red dots). A larger dot indicates more subjects declared for that case. The definition of easy was that the subject felt in good control of the drone, while the opposite falls in one of the other two. The extremely hard was differentiated as a level that the subject feels herself cannot manage the level in reasonable training time.

3.5.2 Online Sessions

During the setup, the subject firstly practiced for at least 10 minutes, and then, piloted the drone once more for personalizing the difficulty without recording signals. Eye-open and eye-close were then recorded for a preliminary calibration of the EEG decoder (see Section 4.1).
Flow of the online sessions, 2 and 3, can be either of the case

<table>
<thead>
<tr>
<th>1st block</th>
<th>2nd block</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 2: Personalize ➔ EEG</td>
<td>Personalize ➔ Manual ➔ Skill Evaluation</td>
</tr>
<tr>
<td>Session 3: Personalize ➔ Manual</td>
<td>Personalize ➔ EEG ➔ Skill Evaluation</td>
</tr>
</tbody>
</table>

A block (condition) in one session has 12 trajectories

Change of Trajectory

Figure 3: Online session. (a) The flow of online sessions, where the personalize and skill evaluation refer to the same task as in Figure S1, with a difference that, skill evaluation itself, does not adjust difficulty levels. Each block is composed of 12 trajectories. The curve below is an example of plotting the difficulty levels during all trajectories against the decision points. The vertical lines indicate the boundaries between trajectories where are the only occasions of decreasing difficulty levels. (b) Example of one trajectory to demonstrate the online interaction in the EEG condition. The blue asterisks are associated to the right y-axis and indicate the events of the increasing level (harder), the subject wanted to increase the level (wants harder), or hit and miss of the decision waypoint. The vertical lines help align the timing of those events, and the lines are extended to the bottom panel in the event of decision waypoints. As shown in the bottom panel, the level is increased if the ratio of Hard class is lower than 0.5, and the ratio is reset at the beginning of a trajectory or at the event of decision waypoints.

Figure 3(a) shows the flow of online sessions. Each session was composed of two personalization processes, one skill evaluation, one EEG block, and one Manual block. The first personalization was done without recording signals as described in the previous paragraph and the skill evaluation is the same task without further tuning the levels. Since skill evaluation is a subset of personalization, please notice that in the later texts, it may also refer to the skill evaluation parts in the personalization processes.

If a subject begins with the EEG condition in the first online session (see Figure 3(a)), The Manual condition will become the first block in the second online session. We divided the subjects into two groups of similar task scores in the first session. The first group had the EEG condition being the first block in the second session and the other group began with Manual condition. The subjects knew the condition before starting each block.

In the EEG condition, the level was increased if the EEG decoder decides as Easy (see Section 4.3 for details of interaction and Figure 3(b) for an illustration). On the other hand, the Manual condition was based on whether the subject pressed the button before the decision point.
As both cases cannot decrease the level in the case of false positive (hard), we unbiasedly lowered the levels when starting a new trajectory. This avoids a dominate class in ground truth for both conditions. Specifically, each condition was conducted in a block of twelve trajectories (see Figure 3(a)). The beginning level of the 1st trajectory was always level 1 while the other trajectories began four levels (as there were eight decisions) lower than the final level of their previous one. For example, the 2nd trajectory began at level 3 if the final level of the 1st trajectory was 7. The curve is an example of concatenating all decision points in the same block. One can see that the level is either increasing or staying at the same level, except for changing trajectories as indicated by the vertical lines.

Each trajectory consisted of 33 waypoints with eight decision points (4th, 8th, ..., and 32nd waypoints). After the decision point, the level either increased or stayed the same. In the case of EEG condition, the operator examined the behavior of the decoder before the 1st trajectory. This also happened before the 5th and 9th trajectories if the level was not moving between Easy and Hard. During the examination adjustment of the bias term of regression (see Section 4) was based on the feedback of the subject, where the subject additionally piloted one time or a few times of an extra and long trajectory. The adjustment did not require the subject to finish the trajectory and the time was made as short as possible.

4 DECODING PERCEIVED DIFFICULTY

4.1 Signal Pre-processing

EEG and EOG signals were downsampled to 256 Hz and casually band-passed between 1 and 40 Hz by a 14th order Butterworth filter. The vertical EOG component was computed by subtracting the sensor between eyebrows by the average of the other two. The horizontal EOG component was derived from the bipolar signal between the two sensors close to canthi.

Out of 39 recordings, P2 was removed twice from the offline or online sessions due to short-circuit with the CMS or DRL electrode. A 20th order spatial low-pass filter, SPHARA [22], was applied to interpolate signals for the removed electrodes and more importantly to reduce high spatial frequency components, likely corresponding to artifacts [22]. After SPHARA, peripheral electrodes were left out of the analysis to reduce the likelihood of muscular contamination, yielding twenty-five channels centered at Cz, namely: F3, F1, Fz, F2, F4, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, CP4, P3, P1, Pz, P2 and P4. Common-average re-referencing was then applied [23]. Potential EEG artifacts due to eye movements were alleviated by Independent Component Analysis (ICA). We used an iterative method to estimate the assumed number of selected independent components. Specifically, we performed RUNICA [24], starting with 15 components, and iteratively searching for the maximum number of components with a proper ICA solution. A proper solution is that the returned weight matrix has no imaginary number and that the maximum and minimum values in the weight matrix are similar, for which we picked 5 as a threshold. If any of the final independent components had a correlation coefficient higher than 0.7 with the vertical or horizontal EOG component, the independent component was then dropped out from the future analysis. On average, 15.8 components were returned by ICA and 1.07 components were then removed during the construction of online decoders.

We took 10log V of Power Spectral Density (PSD) as features, using Thomson’s multitaper algorithm with the time-half bandwidth product being 2. The features were extracted over a two-second sliding window, with a 500ms shift for evaluating the decoder in offline and 125ms for online decoding. This resulted in a 0.5 Hz resolution, and only the power bands between 2 to 28 Hz were extracted, in order to avoid the movement-related slow cortical potentials [25] and too low signal-to-noise ratio. We rejected windows where any EEG time sample had a peak value larger than 50μV after the previous pre-processing. The features were then subtracted by a feature vector from the baseline recording (eye-open and eye-close). The vector was computed by averaging the features across time. In online sessions, we calibrated the features by subtracting the newly recorded baseline.

4.2 Classification Method

Given that the perception of difficulty level is more naturally represented by a continuous value, we first regressed from the log-PSD values to the reported difficulty level (0 to 100, but normalized to 0 and 1) with generalized linear models [26] and elastic net regularization. The chosen distribution and link model were binomial and logistic, respectively, and a grid search was performed for the regularization parameters α = [0.15, 0.5, 1] and λ = [0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20].

The regression model includes a bias term that was shifted in online decoding when the decoder was not operating in a normal range. Since the perceived difficulty was less likely to change instantaneously, we smoothed the output of the regression by a moving average filter. In particular, we averaged the output within the latest 6 seconds to compromise between accuracy and latency. We then classified the smoothed output as Easy (Increase the Level) or Hard (Keep the Level), where the Hard class from now on refers to the collection of both reported Hard and Extremely hard labels unless specified. The classification was performed at each time window using a threshold learned by a Linear Discriminant Analysis (LDA). The LDA was implemented by assuming equal distributions from both classes.

The regression and LDA were trained based on the data of offline session for the online sessions. The regularization parameters of regression were those yielded the best performance during the cross-validation (to be detailed in Section 4.4).

4.3 Online Interaction

Figure 3(b) depicts the principle of the EEG based online interaction in a trajectory, where the blue asterisks are linked to the y-axis on the right and indicate the event of decision waypoints, decisions made (harder), and ground truth given by the subjects (wants harder). The decoder was giving a decision in an 8 Hz rate during online sessions (see the panel with y-axis labeled as Estimated difficulty level). However,
the level can only be increased at a decision waypoint (those vertical lines extended to the bottom panel). Whether to increase depends on the amount of each class within the four waypoints (see the bottom panel with y-axis labeled as Ratio of Hard Class). If the decoder output was dominated by the Easy class, the level increased (harder). Otherwise, the level was kept. The counter of each class was reset to zero when a decision point was met or at the beginning of a trajectory.

4.4 Performance Validation

Both the offline and online decoding performances were assessed by class-balanced accuracy. The $\alpha$ and $\lambda$ hyper-parameters of the regression model for each subject in their online sessions were decided by the best decoding performance in an offline validation. The best performance was decided by maximizing $0.5 \times \text{mean} (\text{across test sets})$ of class-balanced accuracy - $0.5 \times \text{mean} (\text{across classes})$ of standard deviations (across test sets).

The offline validation strategy ideally is four times of leave-one-pair-out cross-validation, where each test set holds a pair of trajectories; one trajectory labeled as Easy while the other as Hard or Extremely Hard, and this allows the assessment of the class-balanced accuracy. However, it is only equivalent to four times of leave-one-pair-out cross-validation if the total amount of Easy trajectories is equal to the total amount of Hard plus Extremely Hard trajectories. Otherwise, it is sixty-four ($4 \times 16$ pairs) times of leave-one-pair-out validation which ensured the highest priority on the least picked trajectories for testing. In this case, some Easy trajectories were chosen less or more times than the Hard plus Extremely Hard trajectories.

For the EEG condition of online sessions, which is referred as closed-loop, two class-balanced accuracies were provided, one evaluated on the entire block and the other is divided into three groups of four trajectories, where the four trajectories of each group always share the same bias term in regression. When evaluating on the entire block, the accuracy was computed based on 96 samples ($8 \times 12$ trajectories), and 32 samples with a group of four trajectories. The ground truth of a decision point was obtained from the button pressing. Once the button is pressed, the current decision point is considered as Easy, and as Hard if not pressed. If a button is pressed but the level stays the same after a decision point, the ground truth of the next decision point was also considered as Easy.

On the other hand, we also evaluated the open-loop cases in the online sessions, where no feedback was provided to the subject. Specifically, the open-loop cases refer to the Manual condition, the skill evaluation part right before the second block, and the skill evaluation in the flow. The case before the first block was not evaluated because it was done during the setup without recording the physiological signals. The purpose was to evaluate the best decoding performance that could be achieved by shifting the bias term in the regression. The evaluation was done by a post-hoc simulation that used the same decoding procedure as in the closed-loop case. We scanned different shifts (add or subtract a value) from the bias term, where the shift ranged from -0.5 to 0.5 with 0.005 as a step.

5 Analysis of Online Behavioral Data

5.1 Task Scores

One question of interest is whether subjects could achieve similar task performance in both conditions. As a result, two-tailed paired t-tests over task scores were conducted between both conditions in each session for a group level analysis ($n=13$ subjects). We further applied a two-tailed paired t-test ($n=12$ trajectories) for each subject in each session. We then applied a Bonferroni-Holm correction to reduce the false positive rate [27]. A perfect decoder should yield similar task scores as the Manual condition. Although the task scores in the personalization or skill evaluation can be measured, they were not compared. The comparison otherwise would not be fair as the functional task difficulty levels would not be the same, because the used eleven levels were the same in all the online sessions; the personalized levels were the sixteen levels used in the twelve trajectories of a condition.

5.2 Skill Curves

Apart from the task scores, skill improvement is also worthy of examination. The sigmoid regression between hit rates and radii (details in Section 5.1) can represent a subject’s skill curve. The skill curves are characterized by several parameters, the two important ones are $x_{50}$ (the x-value has 50% of the range in y response) and slope. Generally speaking, a larger (smaller) $x_{50}$ (slope) represents a worse skill. The comparison was conducted by a two-tailed paired t-test across subjects ($n=13$) either over conditions or between the first and second blocks (to see time effect). The correlation between decoding accuracy and the two parameters were also checked.

The data being used for the skill curves are the hit rates of all levels in each block, where the x-axis represents the 16 levels instead of radii for each block. This is based on the assumption that the same level is mapped to the same point on the functional task difficulty level. The regression was bounded between 0 and 1 for the hit rate ($y$-axis), but the lower bound was relaxed if there is a convergence issue.

In addition to the $x_{50}$ and slope, another index is the area beneath the skill curve. The area was computed by an integral over level 1 to 16 which leads to overall hit rates across all the levels. The larger the area is the better the skill. With the overall hit rates, their differences between both conditions were computed for each subject. The differences were further t-tested ($n=13$) against 0. A significant result indicates that the overall hit rates between both conditions are different. The correlation with decoding accuracy was also computed.

5.3 Final Levels

Assuming the subject’s skill didn’t change drastically and the personalization process worked perfectly, the curve of final levels v.s. trajectories should also be similar between the EEG and Manual conditions, especially, the radii were re-mapped to the same set of functional difficulty levels. In order to compare dissimilarity of the curves, three different indices were defined:
Pearson’s correlation: this gives the similarity of a co-varying trend between two patterns (vectors, \( n = 12 \)), and dissimilarity can be computed as \( (1 - r)/2 \), where \( r \) is the correlation coefficient. Although correlation is a good indicator of similarity, it does not tolerate overshooting but tolerates any magnitude of shifting. We consider \( r > 0.5 \) with significance as two patterns are highly similar.

Mean of difference (MD): it indicates the difference of averaged level between both conditions. It tolerates symmetrical overshoots and undershoots without considering the amount of shooting. We consider MD lower than 1.5 level as two patterns are highly similar.

Mean of absolute difference (MAD): a typical index to compare two curves, it does not tolerate any kind of overshoots or undershoots with a high magnitude. We consider MAD lower than 1.5 level as two patterns are highly similar.

Each index has its pros and cons. Therefore, having at least two indices showing high similarity are convincing to conclude that both curves are similar.

6 RESULTS

6.1 Offline Behavioral Result

The plot of subjective difficulty level in Figure 2 summarizes the reported numerical and descriptive difficulty levels. The curve is the average of all subjects with the shaded area represents the standard deviation. The green, blue, and red dots represent the total amount of Easy, Hard, and Ex. Hard labels, respectively. A larger dot stands for more subjects labeling that class. It can be seen that the designed protocol nicely induced subjective difficulty level in the intended v-shape. For subject-wise result, please refer to Figure S2 in the supplementary material.

6.2 Offline Accuracy Validation

The mean class-balanced accuracy in offline validation across subjects was 76.7% with the standard deviation being 5.1%. These values indicate plausible decoders are available for the online sessions. For detailed result, please refer to the Figure S4 in the supplementary material.

6.3 Online – Decoding Accuracy

6.3.1 Closed-loop

Figure 4 illustrates the online decoding accuracy of the EEG condition for each session and subject. Blue bars show the class-balanced accuracy while green and red bars are class-accuracy for Easy and Hard, respectively. An upward (downward) bar means the accuracy is higher (lower) than the theoretical chance level of a binary problem (50%). The ‘x’ markers are the ratio of Easy samples in the ground truth. If the ratio is 0 or 1, the class-balanced accuracy corresponds to accuracy.

Each sub-figure (session) shows the per-subject result computed with 96 samples (8 decision points \( \times 12 \) trajectories) of the entire EEG block. For both sessions, in total 16 out of 26 (62%) recordings have a class-balanced accuracy higher than the theoretically chance level. The mean class-balanced accuracy in Session 2 (3) across subjects was 56.2% (54.7%) with the standard deviation being 8.6% (11.0%). For accuracies evaluated on four-trajectory groups and the used shifts of bias term during the online decoding, please refer to Section S.II.4 in the supplementary material. The online decoding accuracies were lower than in the offline validation. This can be a typical case for EEG decoders.

6.3.2 Open-loop

Figure 5 plots the highest class-balanced accuracy with the best bias term in the Manual condition (see Figure 5(a)) and the skill evaluations (see Figure 5(b)). In Figure 5(b), each skill evaluation is drawn in a specific color. The upper
panels of Figure 5 report the best accuracy per subject. Similar to Figure 4, an upward bar of accuracy is higher than the chance level, and the one without a bar (s4 in session 2) performed as a chance-level decoder. The lower panels of Figure 5 show the corresponding best shift of bias term. A longer bar regardless of the pointing direction indicates that the decoder is less stable across the offline and the online sessions, while the difference between the two bars in Session 2 and Session 3 indicate the stableness across the online sessions.

Nearly half of the subject had over 60% of accuracy in the Manual condition, and the ranking of accuracies was consistent across sessions (Spearman’s correlation, $r = 0.92$, $p$-value = 0, $n = 13$). The accuracy was therefore highly dependent on the subject. From the panels of shifts, the stableness across the offline and the online sessions was clearly low. Between the online sessions, the average difference of the best shifts was 0.05 with a standard deviation of 0.04, where a shift of 0 between sessions means that the decoder does not need to be tuned again. Given that an averaged inter-session difference is 0.05, this indicates that the best shifts was 0.05 with a standard deviation of 0.04, where a shift of 0 between sessions means that the decoder does not need to be tuned again. Given that an averaged inter-session difference is 0.05, this indicates that the best bias term can be relatively stable across online sessions.

The skill evaluation part did not yield consistent rankings in accuracies between sessions ($r = 0.49$, $p$-value = 0.12, $n = 13$ for the 2nd evaluation and $r = 0.09$, $p$-value = 0.71 for the 3rd one). The average differences of the best shifts were 0.05 with a standard deviation of 0.04, where a shift of 0 between sessions means that the decoder does not need to be tuned again. Given that an averaged inter-session difference is 0.05, this indicates that the best bias term can be relatively stable across online sessions.

The skill evaluation part did not yield consistent rankings in accuracies between sessions ($r = 0.49$, $p$-value = 0.12, $n = 13$ for the 2nd evaluation and $r = 0.09$, $p$-value = 0.71 for the 3rd one). The average differences of the best shifts were 0.05 with a standard deviation of 0.04, where a shift of 0 between sessions means that the decoder does not need to be tuned again. Given that an averaged inter-session difference is 0.05, this indicates that the best bias term can be relatively stable across online sessions.

The skill evaluation part did not yield consistent rankings in accuracies between sessions ($r = 0.49$, $p$-value = 0.12, $n = 13$ for the 2nd evaluation and $r = 0.09$, $p$-value = 0.71 for the 3rd one). The average differences of the best shifts were 0.05 with a standard deviation of 0.04, where a shift of 0 between sessions means that the decoder does not need to be tuned again. Given that an averaged inter-session difference is 0.05, this indicates that the best bias term can be relatively stable across online sessions.

The skill evaluation part did not yield consistent rankings in accuracies between sessions ($r = 0.49$, $p$-value = 0.12, $n = 13$ for the 2nd evaluation and $r = 0.09$, $p$-value = 0.71 for the 3rd one). The average differences of the best shifts were 0.05 with a standard deviation of 0.04, where a shift of 0 between sessions means that the decoder does not need to be tuned again. Given that an averaged inter-session difference is 0.05, this indicates that the best bias term can be relatively stable across online sessions.

6.4 Online – Task Scores

Figure 6 shows the comparison of task scores between the Manual and EEG conditions as boxplots for each online session. Each subject has a red box representing the Manual condition and a blue one for the EEG condition. The horizontal line inside a box is the median and the cross (+) represents the mean value. The colored circle dots are outliers.

Group level analysis using a two-tailed paired t-test ($n=13$) on each session shows there is a large effect of the conditions on the scores, where Manual is significantly better. Session 2 yielded a $p$-value of 0.01 and an effect size of 0.8. Session 3 yielded a $p$-value of 0.005 and an effect size of 0.95.

Two-tailed paired t-tests ($n=12$) with Bonferroni-Holm correction ($m=13$) on each session were also conducted between both conditions as shown in Figure 6 (** $p < 0.05$, *** $p < 0.01$, and **** $p < 0.001$). Although the scores for Manual condition are generally higher than the EEG condition and have a smaller variation, they are not necessarily significantly better. A few cases show that the average score (+ sign) for EEG condition was higher than the Manual one. Specifically, s3 and s10 in session 2 as well as s2 and s5 in session 3 were higher. This implies that using an EEG decoder still has the potential of having a better score.

6.5 Online – Skill Curves

Following Section 5.2, Table 1 lists the results of t-tests for x50 and slope. The condition columns mean that the test was conducted between the Manual and EEG conditions, while the Time columns refer to the result between the first and second presented blocks. Row p is the $p$-value and Power is the t-test statistical power, where a positive value means that Manual (1st block) is higher than the EEG condition.

1. Equivalent to ANOVA since there are only two variables.
Figure 6: Task scores and t-tests (n=12) with Bonferroni-Holm correction (*: p < 0.05, **: p < 0.01, and ***: p < 0.001).

Table 1: T-tests (n=13) on the parameters of skill curves between the conditions and the order of presentation.

<table>
<thead>
<tr>
<th></th>
<th>x50</th>
<th>slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 2</td>
<td>p</td>
<td>power</td>
</tr>
<tr>
<td></td>
<td>-0.15</td>
<td>-0.84</td>
</tr>
<tr>
<td>Session 3</td>
<td>0.34</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 2: Correlations (n=13) between the parameters of skill curves and the decoding accuracy in the EEG condition.

<table>
<thead>
<tr>
<th></th>
<th>x50</th>
<th>slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 2</td>
<td>r</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td>-0.56</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 3: Differences of overall hit rates and the correlation (n=13) between the differences and decoding accuracies.

<table>
<thead>
<tr>
<th></th>
<th>Manual - EEG</th>
<th>1st block - 2nd block</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 2</td>
<td>Avg</td>
<td>Std</td>
</tr>
<tr>
<td>2</td>
<td>0.56</td>
<td>2.72</td>
</tr>
<tr>
<td>3</td>
<td>0.40</td>
<td>2.17</td>
</tr>
</tbody>
</table>

(2nd block) in the condition (Time) column. No significant result was found suggesting there is no strong effect on the condition or time. That means both conditions are similar in terms of the skill curve and the improvement in each block is not biased by the time being executed.

Table 2 shows the correlation between the parameters and the decoding accuracy for the EEG condition. Only the second session has x50 as negatively correlated, which is encouraging as a higher accuracy leads to a lower (better) x50.

On the other hand, Table 3 lists the result of overall hit rates in different sessions for comparing the two conditions and the effect across time. Positive values in the Avg columns indicate that Manual (1st block) is better than the EEG condition (2nd block). No significant correlation was found suggesting that the decoding accuracy did not correlate to either of the checked factors.

6.6 Online – Final Levels

Figure 7 reports the final level of each trajectory for all subjects within each session as histograms. The blue bars stand for the EEG condition with the red ones for the Manual. Their mean values are very close and the majority of the final levels located between level 6 to level 9 regardless of the condition. EEG condition, as expected to be less stable than the Manual condition, shows a higher standard deviation and higher counts on the tails of the distributions.

Figure 8 illustrates the curves representing the final level of each trajectory for three exemplar subjects in both sessions. The red curves stand for the Manual condition while the blue ones are for the EEG condition. Based on all subjects (see Figure S8), the patterns of Manual condition generally show trends that the subjects preferred to stay around a certain level, usually around 8. This behavior is consistent with the behavioral condition of a previous EEG study that adapts the levels of Tetris [15]; as reported by the authors, the subjects preferred to stay around a certain level and as soon as possible. However, after a few more trajectories, our subjects increased a bit the level, probably because subjects felt more confident or more familiar with the control. On the side for the EEG condition, around half of the recordings had similar patterns as the Manual cases.
The correlation, MD (mean difference), and MAD (mean of absolute difference) are also indicated for each subject as texts, where boldfaced values stands for high similarity across subjects are not necessarily the same. (see Figure S 2 in the supplementary material). This indicates that either (1) hit rates did not consistently reflect the subjective feeling, (2) some subjects largely improved their skills, but not others, during the 32 trajectories, or (3) the estimated hit rates were not robust enough as based only on 11 data points. It was not easy to conclude the main factors, but in any case, the data was sufficient to build subject-specific decoders and perform offline validation.

The offline validation yielded promising decoding accuracy. Although the accuracy for some subjects was biased towards one class (see Figure S4 in the supplementary material), the tuning of the bias in regression term during the online sessions seems to help alleviating such situation. Indeed, as long as the best shift of the bias term can be found, the decoding accuracy could still be high.

Many behavioral results in online sessions suggest that the EEG and Manual conditions were similar, even though task scores of Manual condition were significantly better than the EEG condition at the group level. For a subject, the task scores of both conditions could be similar, about two-third statistical tests between conditions did not yield a significant difference in the task scores. The parameters of skill curves also nearly showed no statistical difference. Besides, about half of the cases, the curves of final levels were very similar between both conditions, and nearly every subject had one session with a close pattern. These results suggest that the EEG decoder was able to yield similar results as the self-paced condition. Similar to the study of learning a new arithmetic system [20], the authors also reported that EEG condition yielded similar behavioral results compared to a condition based on user’s behavior, namely the error made. Therefore, the feasibility of using EEG decoder in the interaction loop holds promise.

Another similar, but not comparable, study was made by Faller et al which was also using a simulated drone and have two conditions [10]. There are several differences. In their study, the adaptive process did not change the piloting task, while the difficulty level of our piloting task was changed. They focused on whether down-regulating the arousal state is useful in a high-demand and non-learning task, while we aimed at whether users' cognitive states can replace self-paced decisions in a learning-like setting. Faller et al instructed the subjects to actively down-regulate the arousal state while piloting, and their non-EEG condition utilized sham decoders that were irrelevant to the ground truth. On the contrary, the subjects in our work were not instructed to pay attention to the decoder, and our non-EEG condition was ground truth. In both of our conditions, the subjects should be consistent with the definition of the descriptive label when pressing the button. That means, pressing the button in the online sessions was expected to be equivalent to labeling Easy in the offline session.

An interesting finding in our study is that in rare cases the EEG condition performed better than the Manual one in task scores. One case was s10 in session 2 which had high online accuracy. However, s1 in session 3 also had a high online accuracy but did not surpass the ceiling. One possible explanation is that s10 had lower accuracy in the Hard class than s1. In other words, the decoder tended to make the level harder, forcing the subject to challenge himself by leaving his comfort zone, or s10 was rather conservative on when to increase the level in the Manual condition.

Although the offline validation accuracy was high, it did not necessarily guarantee high online decoding accuracy tested on different days. This cross-day variation might come from fatigue, different background states, different...
The role of decoding accuracy in the interaction is unclear. Ideally, a perfect decoder should give the same or similar result as the Manual condition, and a decoder biased toward one class would lead to a result dominated by level 1 or 16. The correlation to the accuracy in Table 2 and 3, however, only yielded one significant result which supports a higher decoding accuracy leads to a better skill curve. Other correlations tests in Section 6.7 also yielded limited evidence to show that a higher accuracy leads to a better behavioral outcome, scores, skill curve, and high similarity to the Manual condition. As a result, it is hard to draw any conclusion between the decoding accuracy and the behavioral results. One issue was that the average online accuracy was about 20% lower than the offline validation. This means that the online decoding accuracies were generally not much higher than the chance level, and one standard deviation away from the mean was not even higher than 66%. The rather close-to-chance-level accuracy might hinder showing significant results with a small number of samples. Apart from the relatively low accuracy, this indecisive conclusion might also be largely influenced by the personalization of difficulty without many samples. A small number of samples may not give a reliable estimate, but a rather large sample size would take too much time to collect and, more importantly, improved too much the skills of subjects such that one cannot easily observe any learning effect.

We believe that building a robust decoder in the closed-loop condition has a higher priority for better answering the role of decoding accuracy in the future experiment. One working direction is to solve better the cross-session variations. For example, apart from the baseline calibration, we can perform the scan of the best bias term before the actual condition at the cost of much more preparation time and effort that may be less useful for daily usage and lead to lower generalizability. At least, it would be useful to better answer the role of accuracy.

It is worth noting that neural correlates of the subjective difficulty level reported in this paper (see Section S.II.2) are similar to those we found in a previous study where subjects piloted the same drone, but with their left hand (instead of right hand as here) and the difficulty level was modulated by the degree of multi-tasking [30]. Based on common features, the neural correlate of the subjective difficulty level seems to lie in the α band in the right centroparietal hemisphere (C4 and CP4 in particular) and the θ band around Cz.

Our protocol is closer to a realistic task as compared to Faller et al. [10] and we evaluated the decoder with a much shorter latency than Walter et al. [20]. We anticipate that our method may apply to scenarios such as education and gaming. Possible benefits are a shorter learning time and dynamically adapting difficulty levels for action games that are normally fast pace.

Given the competitive behavioral outcomes and accuracy for some subjects, we believe that a certain population may directly benefit from a BMI regulator. Furthermore, as in other BMI applications, in particular related to gaming [31], users may learn to generate better neural correlates of their subjective difficulty level with practice once they experience the benefit of the BMI regulator.

8 Conclusion

The theory of challenge point suggests that leaving the easy zone benefits learning during practice. Following this, the designed online interaction demanded the subjects to leave their comfort (Easy) zone. The closed-loop experiment demonstrated that the behavior results could be similar between using an EEG decoder and self-paced decisions. Subjects might even perform better with an EEG decoder if the decoder is accurate enough and slightly biased toward increasing the difficulty level. However, more positive samples are still needed to draw a decisive conclusion. One potential reason for having lower accuracy might be the duration of a certain level. In the offline validation, each trajectory lasted for around 90 seconds while a level in the online session could be as short as twelve seconds. Apart from improving the accuracy with a scan for the best bias term, another future research direction is examining the applicability of the decoding framework in other tasks with online evaluations, such as in-class education and other kinds of visuomotor responses, e.g. using another hand or feet to control [30]. This will probe whether the decoded cognitive state is task-specific and robust in real interactions.

Acknowledgments

We would like to thank Alexander Cherpollo for his help in the implementation of the simulator and Ruslan Aydahrkanov for his suggestions in designing the protocol. Some figures were drawn with the Gramm MATLAB toolbox [32].

References


Standardization of Neurotechnology for Brain-Machine Interfacing: State of the Art and Recommendations

Research and development of brain-machine interfacing (BMI) systems and related neurotechnologies are at a crucial stage in their history. Progress in sensing technologies, advanced materials, robotics and artificial intelligence provides possibilities that until recently were considered science fiction. Direct neural interfacing with external or virtual devices can usher a new era where merging biological and artificial intelligence will have significant impact in multiple domains.

First and foremost, BMIs are becoming powerful tools to improve our understanding of the brain and nervous system. In turn, this can lead to better therapeutic and assistive approaches to tackle healthcare challenges, as well as new modalities for human-machine interaction that may have transformative effects in many consumer-oriented applications.

Not surprisingly, these technologies have generated remarkable interest and investment from both public and private organizations, including several publicly funded national and regional brain initiatives, as well as the worldwide creation of a large number of neurotechnology enterprises. Some projections expect the neurotechnology market to reach a valuation of USD 19 billion by the end of 2026 [1].

Despite their promise, BMI may be on the cusp of the hype curve, facing increasing pressure to demonstrate concrete value to users. In addition to the numerous technical challenges inherent to developing safe, efficacious, and reliable solutions, researchers and developers face the complex human-centered challenges of discerning which data and use cases provide the most value to which users and organizations.

The development and commercialization of BMI systems require researchers, clinicians, manufacturers, and regulatory bodies to ensure that these devices comply with well-defined safety and effectiveness criteria. BMI systems typically require integration of multiple modules comprising measurement and analysis of neural activity, and provision of feedback to the user through various means, such as visual displays, virtual reality systems, haptic interfaces, and exoskeleton. The scarcity of specific BMI and broader neurotechnological standards hinders the design of new devices for interoperability and regulatory compliance, thus posing a barrier to widespread user access (industrial, clinical, and consumer) and potential benefit.

It is thus imperative for the BMI community to have a good understanding of the current state of the standards in the field, as well as the main gaps that need to be addressed.

For this reason, the IEEE Standards Association (IEEE-SA), IEEE Engineering in Medicine and Biology Society (EMBS)’s Technical Committee on Standards, and IEEE Brain Initiative initiated an Industry Connections Activity (ICA) on the topic of Neurotechnologies for Brain-Machine Interfacing (NT-BMI; IC17-007) [2]. This initiative is dedicated to evaluating existing standards and best practices for BMI system design and usage, as well as to identifying priority areas for new standards. The NT-BMI established a multi-stakeholder group, comprising experts and representatives from academia, industry, and regulatory agencies worldwide. In February 2020, we released an IEEE Standards Roadmap [3] providing a comprehensive overview of the current practices and future requirements for NT-BMI standardization. This activity has also spawned three Standards Working Groups: IEEE P2725.1: Standard for Microwave Structural, Vascular or Functional Medical Imaging Device Safety [4, p. 1], IEEE P2794: Reporting Standards for in vivo Neural Interface Research (RSNIR) [5], and IEEE P2731: Standard for a Unified Terminology for Brain-Computer Interfaces [6].

BMI systems typically integrate multiple elements or components, often comprising technologies at different levels of maturity. Available standards may thus vary considerably across constituent elements. Since most BMIs place the “user-in-the-loop,” such standards should address the end user’s needs, attention (engagement) and intention, including user instructions. To reflect the nature of BMIs as ‘complex systems of systems,’ the NT-BMI Standards Roadmap is structured in five functional areas identified by the NT-BMI Group: (1) sensor technology, (2) end effectors, (3) data representation, storage & sharing, (4) user needs, and (5) performance assessment & benchmarking. This editorial and accompanying Emerging Topics papers in this journal present and discuss the main findings and recommendations of the NT-BMI and related working groups.

BMI sensor technologies: encompass a broad spectrum of transducer types, including both invasive and non-invasive modalities. They range from well-established and widely used techniques such as electroencephalography to emerging approaches like microwave and ultrasound imaging, stentrodes, neural lace, and neural dust.

Among the five functional areas, sensing technologies are arguably the area with the highest level of standardization. Nonetheless, there is no established standard for time synchronization among different systems and modules, since the interfaces and ports to those systems vary widely. The NT-BMI Group also recommends that consumer-grade sensors comply
with safety and performance standards consistent with clinical device requirements, given the prevailing trend towards use of consumer device data for health and wellness applications [7].

*End effector systems for BMIs* include actuators, virtual or physical devices, and feedback mechanisms. They can be broadly categorized into exoskeletal devices, prosthetic devices, virtual/augmented reality interfaces, and neurostimulation devices (peripheral, spinal, transcranial, and intra-cranial).

Priorities for standardization in this functional area include data communication protocols between the end-effector and other BMI elements, shared control strategies and architectures, and unification of terminologies. The first paper of this series, “A Roadmap Towards Standards for Neureally Controlled End Effectors” [8] provides more detailed information on this topic.

*Data Representation, Storage, & Sharing:* There have been a variety of efforts to define data formats for various biosignals used in BMI systems, in the forms of file format specifications, standards, software frameworks, and initiative groups. Nonetheless, efficient data storage and secure interoperability has emerged as the ‘need of the hour’ for standardization – in particular, specific to closed loop applications. Similar to other highly-sensitive-data-based applications, requirements for portability, interoperability, and privacy are essential for viable BMIs and associated systems. To this end, the data standards now being developed by IEEE P2933 WG (“TIPPS for Clinical IoT”) may provide a useful framework for BMIs [9].

IEEE P2731 WG is also working on defining the information that should be stored into data files to allow automatic processing of BMI signals without the need to access additional resources (e.g., scientific papers or other documents), which is time-consuming and requires human intervention.

*User Needs:* The specification of device users, use cases, and the fulfillment of user needs remain foundations of the user-centered design (UCD) process for both medical and consumer devices. Indeed, UCD processes (including human factors/usability engineering: HFE/UE) have been shown to yield significant downstream benefits in product development life cycles, including higher user satisfaction, better product adoption, reduced net development costs, and early insight regarding future products and markets [10]. While usability evaluation is a required element of risk management for medical devices and there exist high-level standards defining HFE/UE frameworks [11], [12], the development and maintenance of HFE/UE processes for specific devices remains the resource-intensive responsibility of developers.

To promote the effective, efficient identification and fulfillment of user needs, NT-BMI standardization efforts should thus develop additional HFE/UE standards that complement existing frameworks by defining technology-specific methodologies and quality metrics, in a manner adaptable to different users and use cases [13]. Such standards will improve the rigor of neurotech R&D, the quality of resulting devices, and will reduce the time and resources required for clinical validation and commercialization.

*Specification, Performance Assessment & Benchmarking:* have been identified as additional clear priorities for standardization. Importantly, these protocols and metrics should extend beyond the separate evaluation of individual sub-systems/components and allow assessment of the entire BMI system during closed-loop operation under intended use conditions. The lack of consensus terminology, metrics, and reporting criteria to this end hinder the assessment and comparison of different systems used for related applications. Accordingly, the second paper of the present NT-BMI series formulates a “Functional Model for Unified Brain-Computer Interface Terminology” [14]. In complement, the third paper in the series presents a set of “Preliminary Minimum Reporting Requirements for in-vivo Neural Interface Research” for implantable neural interfaces [15].

By integrating standardized benchmarking protocols and metrics, commonly agreed-upon terminology, and comprehensive scientific reporting guidelines, the NT-BMI initiative seeks to cultivate an ecosystem of increased information interoperability spanning the fields of neuroscience, neurotechnology, and neural rehabilitation. By enabling more rigorous psychometric investigations, this interoperability will in turn promote more robust fulfillment of user needs and better alignment of NT-BMI to serve collective human health and wellbeing. To this end, such technological standards must complement broader initiatives on the ethical and responsible development of technology such as the IEEE NeuroEthics Framework [16], the Ethically Aligned Design guidance [17], and the OECD recommendations for Neurotechnology Enterprises [18].

*General recommendations:* Beyond the specific functional areas, the NT-BMI Standards Roadmap has also distilled the following general recommendation: (1) Efforts should be invested to educate the BMI R&D community on the benefits of standardization with respect to technological design, quality of research, and the ultimate potential for clinical and commercial development. Accordingly, the standards development process should incorporate the perspectives and interests of all neurotechnology stakeholders – including researchers, clinicians, developers, regulatory and scientific reviewers, end users, etc. – via active community engagement by the NT-BMI Group and related initiatives; (2) BMI safety, security and privacy appear as top priorities for standardization. BMI-specific standards in this domain should build on existing principles, standards, and regulatory guidelines for medical and information technologies; (3) Existing efforts to improve scientific reproducibility and open science can be leveraged to establish and consolidate standards for data sharing and reporting on neurotechnology developments; (4) Stakeholders should consider defining complementary and modular standards that promote interoperability, translation, and scaling between consumer and clinical applications; (5) It is important to envision and develop a flexible yet consistent neurotechnology standardization ecosystem that harmonizes community-established best practices, soft law, ethics, international consensus standards, research reporting guidelines, and government regulation. BMI-specific standards should be aligned with existing and emerging standards and regulatory frameworks to address ethical, legal, and societal implications of emerging technologies.

To conclude, it is important to recognize that the scientific and technological foundations for BMI are in perpetual evolution.
Hence, the standardization priorities and recommendations identified herein should be re-evaluated through a continual dialog among all stakeholders. The work presented in the NT-BMI Standards Roadmap and this series of papers is thus intended as an invitation to continue this dialog. Those wishing to collaborate or provide feedback are encouraged to contact the corresponding author(s) of interest.

RICARDO CHAVARRIAGA  
Chair, IEEE SA IC Activity - Neurotechnologies for Brain-Machine Interfacing (IC17-007)  
Zurich University of Applied Sciences, ZHAW  
Winterthur, Switzerland  
r_chavarriga@ieee.org

CAROLE CAREY  
Chair, IEEE EMB Standards Committee  
Engineering in Medicine and Biology Society  
c.carey@ieee.org

JOSE LUIS CONTRERAS-VIDAL  
FIEEE, FAIME  
Co-Chair, IEEE SA IC Activity - Neurotechnologies for Brain-Machine Interfacing (IC17-007)  
NSF IUCRC BRAIN  
University of Houston, Houston, Texas  
jcontreras-vidal@uh.edu

ZACH MCKINNEY  
Chair, IEEE P2794 Standards Working Group – Reporting Standards for in vivo Neural Interface Research (RSNIR)  
The BioRobotics Institute; European Ctr of Excellence in Robotics & AI, Scuola Superiore Sant’Anna, Pisa, Italy  
z.mckinney@ieee.org

LUIGI BIANCHI  
Chair, IEEE P2731 Standards Working Group – Unified Terminology for Brain-Computer Interfaces  
Civil Engineering and Computer Science Engineering,  
“Tor Vergata” University of Rome, Italy  
luigi.bianchi@uniroma2.it

ACKNOWLEDGMENT

The authors gratefully acknowledge the IEEE Engineering in Medicine and Biology Society (EMBS)’s Technical Committee on Standards, the IEEE Brain Initiative, and the IEEE-SA Industry Connections Activity (ICA) on NT-BMI. The work of Jose Luis Contreras-Vidal was supported by the NSF IUCRC BRAIN award #1650536 and NSF PFI award #1827769. Authors would like to thank the contributions of all members of the NT-BMI Industry Connections Activity as well as participants in the different workshops and panel discussions organized by this activity.

REFERENCES


Closed-loop EEG study on visual recognition during driving

Ruslan Aydarkhanov, Marija Ušćumlić, Ricardo Chavarriaga, Lucian Gheorghe and José del R Millán

1 Medical Image Processing Laboratory, Center for Neuroprosthetics, Interschool Institute of Bioengineering, École Polytechnique Fédrale de Lausanne (EPFL), Campus Biotech H4, 1202 Geneva, Switzerland
2 Nissan International SA, La Pièce 12, 3180 Rolle, Switzerland
3 École Polytechnique Fédrale de Lausanne (EPFL), Campus Biotech H4, 1202 Geneva, Switzerland
4 École Polytechnique Fédrale de Lausanne (EPFL), Campus Biotech H4, 1202 Geneva, Switzerland
5 ZHAW Datalab, Zurich University of Applied Sciences, Building TD, Winterthur 8401, Switzerland
6 Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, United States of America
7 Department of Neurology, The University of Texas at Austin, Austin, TX 78712, United States of America
8 Imaging-Assisted Neurorehabilitation Lab, Department of Clinical Neurosciences, Faculty of Medicine, University of Geneva, Av. de Beau-Séjour 26, Geneva 1211, Switzerland

E-mail: ruslan.aydarkhanov@alumni.epfl.ch

Keywords: brain–computer interfaces, electroencephalography, eye tracking, driving, visual recognition

Abstract

Objective. In contrast to the classical visual brain–computer interface (BCI) paradigms, which adhere to a rigid trial structure and restricted user behavior, electroencephalogram (EEG)-based visual recognition decoding during our daily activities remains challenging. The objective of this study is to explore the feasibility of decoding the EEG signature of visual recognition in experimental conditions promoting our natural ocular behavior when interacting with our dynamic environment. Approach. In our experiment, subjects visually search for a target object among suddenly appearing objects in the environment while driving a car-simulator. Given that subjects exhibit an unconstrained overt visual behavior, we based our study on eye fixation-related potentials (EFRPs). We report on gaze behavior and single-trial EFRP decoding performance (fixations on visually similar target vs. non-target objects). In addition, we demonstrate the application of our approach in a closed-loop BCI setup. Main results. To identify the target out of four symbol types along a road segment, the BCI system integrated decoding probabilities of multiple EFRP and achieved the average online accuracy of 0.37 ± 0.06 (12 subjects), statistically significantly above the chance level. Using the acquired data, we performed a comparative study of classification algorithms (discriminating target vs. non-target) and feature spaces in a simulated online scenario. The EEG approaches yielded similar moderate performances of at most 0.6 AUC, yet statistically significantly above the chance level. In addition, the gaze duration (dwell time) appears to be an additional informative feature in this context. Significance. These results show that visual recognition of sudden events can be decoded during active driving. Therefore, this study lays a foundation for assistive and recommender systems based on the driver’s brain signals.

1. Introduction

Brain–computer interfaces (BCIs), especially those based on non-invasive electroencephalogram (EEG) signals, are not only proving their value as assistive tools for people with disabilities [1–6] and potential rehabilitation tools for neurological patients [7–10] but also open opportunities to augment interaction for users without disabilities [11–13]. In this latter respect, a promising possibility is to decode neural correlates of perceptual and cognitive processes while people overtly interact with real-world environments [14–16]. In this work, we explore the feasibility of decoding visual recognition from EEG in experimental conditions mimicking daily activities, which involve overt visual search. In particular, we performed a study aiming at EEG decoding of visual recognition during driving. Here the driver is...
primarily engaged in controlling a car simulator, actively exploring the visual environment to complete a visual search task.

Previous works have reported EEG signatures of visual recognition in simple, well-controlled experimental setups [15, 17–19]. The classical restrictive conditions require the stimuli to be static and undergo sharp transitions (e.g., flashing) and subjects to sit still in front of a screen [18]. Such experiments often use an oddball paradigm where subjects had to recognize rare target stimuli in a sequence. In these cases the recognition process is reflected in EEG as the well-known P300 component of event-related potential (ERP). P300 is a positive deflection in the parietal region, which occurs typically between 250 and 500 ms after the stimulus onset [20]. Various BCI applications such as P300-based speller have successfully used this signal. With a typing speed of 10 characters per minute [21], such spellers’ home use may improve the quality of life of people with severe motor disabilities, such as ALS [22–24]. For people without motor impairments, however, this setup and typing performance does not bring much value.

Successful decoding of visual recognition in free viewing tasks can provide novel opportunities for BCI application for healthy users. Car drivers are exposed to a richer environment and constantly interact with it by controlling the car, visually exploring the surroundings, and planning upcoming actions. A potential application could use BCI to decode the driver’s recognition of road signs of a particular target category, e.g., parking. The driver’s interest in parking signs decoded from the EEG will allow to provide timely recommendations for the nearest parking.

Visual search tasks in free viewing conditions require us to fixate static or track moving objects to perceive them in all details. Fixations evoke a type of ERP called eye fixation related potentials (EFRPs) as opposed to external events in more classical ERP designs. Early EEG deflections in EFRP reflect the processing of low-level features of visual input projected from the retina to the occipital lobe and are manifested mainly in high-amplitude of the P100 component, also referred to as lambda component [25]. When fixating at a target object in free viewing visual search tasks, later EEG deflections in EFRP resemble the P300 component in the oddball paradigm [26–28].

Visual stimuli used in previous ERP and EFRP studies range from simple geometric shapes in static scenes to static natural images or dynamic scenes with geometric shapes [29, 30]. Only a few attempts on decoding ERP or EFRP in videos or virtual reality simulations have been reported, e.g. [15, 19]. However, the experimental conditions in those studies did not fully reflect the real-world dynamics. For instance, in human action recognition from a cartoon animation, the video playback was sped up to limit the time of the recognition process [19]. The classification performance of Target vs. Non-Target event reached average AUC > 0.8. In another study involving maze navigation, subjects experienced fast autonomous driving and had to press a button whenever the car in front braked suddenly [15]. The braking task required constant monitoring of the car in front, leading to a brief visual attention of the stimuli. It resembled the attentional load of real driving, however, unnaturally high speed and lack of full car control limit the transfer of the results to real driving. The classification performance based on a single EFRP was around 0.65 AUC on average across subjects.

In this paper, we report a study on visual recognition decoding during driving in a car simulator based on the EFRPs and gaze behavior. In our study, 13 subjects visually attended signboards along the roads while primarily engaged in active driving through an urban environment with a natural speed of ∼80 km h⁻¹. We considered the sudden appearance of the task-relevant visual content (the boards) in the driving environment. Furthermore, the boards were invisible until the driver got close enough, then they popped up at random locations above the sidewalks. Such a stimulus presentation ensured that drivers can recognize the content once they gaze at a board. This setting was chosen to prevent two challenges in decoding visual recognition due to fixations on distant and poorly recognizable boards: (1) to locate the recognition timing within long attendance, (2) to select one of multiple returning fixations when the recognition occurs. The latter impedes the collection of clean data while the former challenges the EFRP decoding itself. Future studies should complement the findings presented here to consider these situations.

The experiment consisted of two phases: offline and online. The goal of the offline phase was two-fold: (1) to analyze ocular and neural correlates of visual recognition and (2) to gather data to train a single-trial classifier that differentiates between attending boards with the target and non-target symbols to be used in the BCI application (the online phase). Then we conducted the online phase, which allows us to examine the single-trial classifier performance in a closed-loop setting and, as a major objective, to verify the corresponding BCI application’s feasibility for detecting the drivers’ interest along a road segment.

Our protocol closely resembles the real-life scenario of traffic sign recognition on the roads. In the application, the BCI system that decodes the cognitive response on the target sign recognition can be integrated into the advanced driver-assistance system (ADAS). In this case, the target is naturally chosen by the driver according to the current driving situation. The accurate decoding of the selection of target traffic signs from the driver’s brainwaves can shed light on the driver’s intentions and goals, from which the ADAS system could derive useful recommendations and assistance.
2. Experimental setup and protocol

2.1. Data collection

Thirteen volunteers (age of 28 ± 6, 3 female, 2 left-handed) with normal or corrected-to-normal vision (8 with contact lenses) participated in the study. One of the participants had a consistently poor eye tracker calibration quality, which led to poor fixation extraction. This participant was excluded from the analyses. The study was approved by the ethical committees of the Cantons of Vaud and Geneva, Switzerland (Commission cantonale d’éthique de la recherche sur l’être humain, study no. PB_2017-00295), and all the participants provided written consent. The experiment lasted 3 h, including the set up (∼1 h), the offline phase (∼45 min), one pause period (∼30 min), and the online phase (∼45 min). The extended pause period was necessary to process all the data from the offline phase and to train the classifier for the online phase. A run represents a ride through the city.

The offline phase consisted of 3 runs, whereas the online phase could comprise 3–5 runs depending on the available time and the subject’s fatigue. One run included 20 road segments with 12 boards each, resulting in 240 boards per run. Before each run, the subjects were asked to move their eyes up-down and left-right for 1 min in order to collect the data for eye movement artifact removal.

The EEG was acquired with the Biosemi ActiveTwo system (Biosemi, the Netherlands) with 64 electrodes at a 2 kHz sampling rate. Additionally, we recorded three EOG channels to collect the eye movement data: two electrodes next to the outer canthi of the eyes and one above the nasion. The real-time processing of EEG in the online phase was done on a dedicated computer.

The eye gaze was recorded with the SMI RED (SensoMotoric Instruments, Teltow, Germany) eye tracking system with a sampling rate of 120 Hz. The chair and eye tracker positions were adjusted for each subject. The eye tracker was calibrated with 13 points once after the EEG setup and before the beginning of the experiment.

The driving simulator logged the car location in the virtual environment, the controllers’ state (gas and brake pedals, steering wheel, buttons on the wheel), and the 2D position of boards on the screen at a sampling rate of 256 Hz. In order to synchronize the data acquisition on three separate machines (EEG, eye tracking, and driving simulator) at different sampling rates, a square pulse of 4 Hz was generated by the driving simulator and sent to the eye tracker through TCP connection and to BioSemi through the parallel port.

2.2. Driving simulator

In our study, we used the driving simulator previously used in [11, 12, 31]. It allows for an immersive driving experience by utilizing a real Nissan driving chair with a steering wheel and two pedals (gas and brake). The simulated car has an automatic transmission, which excludes actions for the manual gear shift. The visual input is provided with three 3D monitors, which create multiple renders for different angles. The virtual environment is implemented using the open-source driving simulator project VDrift. The environment resembles a regular grid city with static objects, i.e. buildings, traffic lights, fields. The task-related items include direction indications on the road, target cues, boards with symbols, and finish lines (figure 1).

2.3. Tasks

The experimental session had two phases: offline and online. In both phases, the subject is instructed to drive through the city (without stopping) while following indications left or right turn at the crossings (figure 1). At the beginning of a road segment, a symbol (the cue) is depicted on the ground. Subjects must memorize and search for it among the boards appearing through the road segment. The data collected in the offline phase allows us to train a single-trial classifier that differentiates between attending the target and non-target symbols. The online phase illustrates how this classifier can be used in the closed-loop BCI application for detecting the driver’s symbol of interest within a road segment. In particular, the single-trial classifier is used to decode the driver’s brain responses to multiple instances of symbols along a road segment. The BCI system infers the target symbol based on this accumulated evidence and shows it to the driver as feedback at the end of a road segment.

2.3.1. Offline phase

Subjects are asked to search and count the number of boards with the target symbol along a road segment. To keep subjects engaged in the task, they have to report the number of targets at the end of a road segment. After crossing the finish line, the subject presses a button located at the steering wheel as many times as the number of target boards he/she counted along the current road.

2.3.2. Online phase

Subjects are instructed to search for boards with the target symbol and pay attention to the feedback at the end of a road segment—after crossing the finish line, the inferred target symbol is rendered at the bottom of the screen as a 2D object (figure 1). If the target identification was correct, a green circle is also shown. Otherwise, the circle color is red. The online phase is a closed-loop interaction because the subjects perceive the decoder result shortly after looking at the stimuli, within 15 s after the first board on the road.

In both phases, we included empty road segments (without boards and cues) where subjects have no additional tasks except driving, allowing them to rest from the visual recognition task. In total, a quarter of the road segments were empty.
2.4. Stimuli presentation

2.4.1. Stimuli visibility

The board presentation is carefully adjusted to guide the visual behavior of subjects. First of all, boards are invisible unless the driver approaches them close enough to make them pop up. Their positions are generated using the following rules. The boards appear along the road at an equal distance between them, randomly on either side of the road above the sidewalks, with a maximum of two boards on the same side in a row (e.g. two boards on the left in Figure 1). The number of boards on the left and right sides is balanced. The pop-up distance was greater than the distance between the boards along the road, creating a time overlap in the presence of multiple boards on the same side. For this reason, the predefined boards’ positions (horizontal and vertical coordinates) were chosen to avoid the overlap for the driver view.

The car’s maximum speed was limited to 80 km h⁻¹ to ensure that all the targets could be attended. The subjects were allowed to slow down if necessary to attend all the boards and count the targets. Nonetheless, all the subjects practiced until they felt comfortable completing the recognition task at the maximum speed during the EEG setup. Due to nearly constant speed and regular placement of the boards, there was an interval of about 900 ms between board onsets.

2.4.2. Stimuli content

In order to link the perception of the symbols on the board with the eye fixations, the recognition by peripheral vision must be avoided. Therefore, the target and non-target symbols were similar and surrounded by the # character to create a crowding effect and force the foveating on the main symbol as done in a previous study [32]. Additionally, we added a bright red border around the board, similar to the traffic signs, to create a contrast with the environment and facilitate their identification (Figure 2).

2.4.2.1. Symbols in the offline phase

In the offline phase, one of the two symbols were depicted on each board: ∃ and ∄. One of them was randomly chosen as a target per road segment and was presented as the cue at the beginning of the road. There were 2–4 targets out of 12 boards on each road segment.

2.4.2.2. Symbols in the online phase

In the online scenario that corresponds to the BCI application, we included four different symbols. It gets symbol appearance is still a rare event. Therefore, we used all four boards.

The offline and online phases of our experiment have been summarized in Table 1.

![Figure 1. Visualization of the experimental setup and protocol. Top left: The experimental setup with the driving simulator (car seat, steering wheel, pedals and 3 3D screens), Eye-tracker and subject with EEG cap. Middle left: A screenshot from the beginning of the road with the cue on the floor and the first board on the left-hand side. Bottom left: A screenshot (in the online phase) showing the feedback after the finish line for the correct target symbol detection—green circle and the 2D board of the inferred target symbol. Right: The schematic drawing of a road segment with the target cue, boards, indications left/right turn at the crossings and finish line. Note that the boards are invisible until they are approached to a close distance.](image-url)
2.4.2.2. Symbols in the online phase
In the online scenario that corresponds to the BCI application, we included four different symbols. It is done to mimic the diversity of the natural visual environment (e.g. traffic signs) and to estimate the generalization of the training with the oddball paradigm. Since the BCI application detects the symbol drivers are interested in along a road segment, four symbols allow an equal occurrence rate of symbols and equal evidence accumulation, while the target symbol appearance is still a rare event. There were between 2 and 4 boards of each symbol resulting in 12 boards on the road. Only one symbol was a target on each road segment. Therefore, the proportion of the targets was 0.25 on average per road section in a run as in the offline phase.

In order to infer the target symbol at the end of the road, each EFRP associated with the symbol attendance is assigned a probability to belong to the target class. At the end of the road segment, the probabilities of all EFRPs are aggregated symbol-wise by averaging. The symbol with the highest average probability of being a target class is selected as an identified target.

The major protocol differences between offline and online phases are summarized in Table 1.

| Table 1. The protocol differences between offline and online phases. |
|--------------------------|-------------------|
| Offline                  | Online            |
| Types of boards          | 2                 | 4                 |
| Task                     | Count silently,   | Count silently,   |
|                         | button press at   | observe the       |
|                         | the end           | feedback          |

3. Methods
The offline and online phases of our experiment have different objectives. As earlier explained, the offline phase is necessary for collecting good quality data for training a classifier that differentiates between attending target and non-target objects to be used in the closed-loop BCI in the online phase. However, it also allows us to study EEG correlates of visual recognition and perform a comparative study of different classification algorithms. The online phase allows the estimation of the closed-loop BCI application performance. In the online phase, we apply a chosen classifier trained on all the offline data.

To make a comparative analysis of different decoding (classifications) approaches under online conditions, we performed a simulated online analysis. Namely, we trained different classifiers on all the offline data and evaluated their performance on the data recorded in the online phase.

The data processing for offline and online phases as well as for simulated online analysis is summarized in figure 3.

3.1. Fixation extraction and analysis
There exist numerous methods to extract eye movement events from the eye gaze direction [33, 34]. We used a robust and accurate method of identification by 2-means clustering (I2MC) [35] for the offline phase data. In the online phase and simulated online analysis, we used identification by dispersion-threshold (IDT), a simpler and easy-to-implement method for real-time processing [36].

3.1.1. Offline phase
We relied on the I2MC implementation provided by the authors of the method. The main idea behind this method is to find the transition between two consecutive fixations by applying 2-mean clustering in a sliding window manner. During a fixation, the eyes do not move, so if we can clearly detect two clusters of gaze direction within the sliding window, it means that they correspond to two fixations. This method is more precise and robust to noisy outliers, leading to a higher quality training dataset. We used the values suggested by the authors for most of the numerous parameters. The major changes include the minimum duration of the fixation adjusted to match the value used for the IDT method (100 ms) and a removal of the downsampling step due to a lowering sampling rate than in the original work.

3.1.2. Online phase
Since the available implementation of the I2MC method cannot be applied in real time to extract eye...
fixations, we used the IDT supplied with our eye-tracking system. Fixation in IDT is extracted when the signals lie within the dispersion thresholds for at least a minimum fixation duration. It requires two parameters: we used 100 ms for the minimum fixation duration and 200 pixels for the maximum dispersion.

3.1.3. Simulated online phase
The same procedure for fixation extraction is applied as in the online phase.

3.1.4. Fixation analysis
The P300-like components of cognitive response are stronger when the stimulus is perceived and recognized for the first time [30]. We assume that subjects categorized the symbol at the first attendance, so we use only the first fixations on the boards for our analysis. As mentioned in section 2.4, the popping up effect and the boards’ design ensure that recognition occurs after fixating on the symbol.

The visual input during the task is dynamic. Due to driving through the virtual environment, the objects, including the boards, are also moving on the screen. To judge whether the board contains the target symbol, the subjects had to fixate it and follow it by gaze until they made a decision. So we assume that most of the board attendances are done with smooth pursuit rather than fixations. However, the onset of the first fixation on a board will coincide with the onset of smooth pursuit. To the best of our knowledge, there is no available algorithm for efficient extraction of smooth pursuit for eye movement data sampled at 120 Hz (the upper sampling rate limit of the eye-tracker model that we used). The only consequence of extracting fixation from smooth pursuit is that a single smooth pursuit may be oversegmented into multiple fixations. For the sake of our analysis, we do not need to differentiate between fixations and smooth pursuit movements. The onset of the first fixation on a board will coincide with the onset of smooth pursuit.

Each detected fixation was assigned to one of the three categories: a target board fixation, a non-board fixation. Considering that (i) the visual span (the angular span) within which our vision is sharp enough to perform actions whereas the text without boxes represents data.
we estimate the probability of fixating eyes on the board’s center according to a normal distribution. After averaging log-probabilities across the fixation time window, we apply a hard threshold to assign the fixation to a board or a non-board category. This procedure was implemented in the analysis of the data. The real-time assignment of fixations on boards in the online phase was modified for computational efficiency. It was based on the first 100 ms after fixation onset and used on the average fixation direction within 100 ms.

For the gaze analysis, we estimate the dwell time—the uninterrupted time which the driver spent looking at a board. The dwells were created by merging all the fixations on the same board with the interfixation interval between them below 50 ms. We used the IDT method to extract fixations for the sake of a valid comparison between offline and online phases in gaze analysis.

### 3.2. EEG processing

#### 3.2.1. EEG processing of offline data

All the EEG and EOG were downsampled from 2 kHz to 256 Hz using FIR anti-aliasing lowpass filter and further filtered with a Butterworth band-pass filter of order 4 within the band \([1, 10]\) Hz in a forward-backward way (to eliminate phase distortion). The filtering range is characteristic for the P300 component [38]. Due to the low conductivity of the skull and the skin, the EEG signal is spatially smoothed, so a high contrast between nearby channels results from noise and movement artifacts. We remove this noise by keeping only low spatial frequency components after decomposition EEG with SPHARA [39]. Horizontal and vertical components of eye movement were estimated, which allowed removing the eye movement artifacts from EEG using multiple regression as in [40] using only horizontal and vertical components as well as the intercept. The coefficients of multiple regression were estimated from the 1 min session of eye movements before the corresponding run. Then the signal is spatially filtered with common-average-reference (CAR).

#### 3.2.2. EEG processing in the simulated online analysis

The EEG was band-pass filtered by the same filters as in the offline phase. Besides, the eye movement artifacts removal is done using the multiple regression model obtained from the calibration data recorded in the offline phase.

#### 3.2.3. EEG processing in the online analysis

The closed-loop interaction in the online phase required real-time processing. SMI system provides real-time eye fixation detection through the IDT method. The fixations were buffered by a parallel process within the driving simulator, matched with the boards, and a trigger was sent to the BioSemi system 3 s after the onset of each fixation on a board. We chose 3 s delay because we applied a non-causal filter on EEG data and needed to diminish the edge effect of the backward filter pass. EEG processing was identical to the offline procedure except for two steps:

- the spectral filtering was done on a 5 s buffer of data, around \([-2, 3]\) s in reference to the fixation onset;
- the eye movement artifacts were removed based on the multiple regression coefficients trained with the eye movement data from the offline phase.

### 3.3. Decoding approaches

To decode visual target recognition from EEG signals, we solve the binary classification problem to classify EFRP epochs into Target vs. Non-target. We extracted the EEG epochs around the fixation onset—more precisely, the EEG activity within the time interval from 200 to 1000 ms following the fixation onset, which is equal to 205 time points. Given the 64 channel montage, a single EEG epoch has a dimensionality of \(x \in \mathbb{R}^{205 \times 64}\) or \(x \in \mathbb{R}^{13120}\) after vectorization. In addition, we extracted dwell time as eye-gaze based features to detect visual recognition (see section 3.1.4). We investigate and compare different decoding approaches, which consist of various feature sets and classifiers.

#### 3.3.1. Offline and simulated online decoding

The following combination of features and classifiers were applied in the offline and simulated online decoding:

- **PLR-Waveform.** Penalized logistic regression (PLR) trained on waveform features (i.e. the signal’s value at each channel and time point) after reducing the dimensionality with PCA. Only the components which explain 90% of the variance are kept (transformed feature vector \(z \in \mathbb{R}^{86 \pm 11}\)).
- **PLR-Dwell.** PLR trained on dwell time on the boards \((z \in \mathbb{R}^{1})\).
- **PLR-Combined.** PLR trained on the combination of waveform features with dwell time. We concatenate the dwell time and EEG waveform features after applying PCA to keep 95% of variance \((z \in \mathbb{R}^{87 \pm 11})\).
- **RF-Waveform.** Random forest trained on waveform features. We use 100 decision trees and with a maximum depth of 5 \((z \in \mathbb{R}^{13120})\).
- **PLR-Riemann.** PLR trained on Riemannian features from simple EEG epochs. To build Riemannian features, we estimate a spatial covariance matrix with shrinkage and project it to the tangent space according to the classical Riemannian geometry on SPD matrices [41]. We subselected eight channels based on the mean Fisher score across the epoch \((z \in \mathbb{R}^{86})\).
- **PLR-Riemann+.** PLR trained on Riemannian features from augmented epochs. Before computing
the covariance matrix, we augment the epoch with the averaged ERP for each class (target and non-target). Otherwise, it is identical to the previous approach \((z \in \mathbb{R}^{300})\).

Since PLR is a linear regularized classifier, we standardize all the features to z-score when using PLR. For PLR, we applied elastic net regularization with a major contribution from \(\ell^1\)-norm \((1000 times greater than \(\ell^2\)-norm; \(\lambda_2 \gg \lambda_1\)).

### 3.3.2. Online decoding

The preliminary tests on two subjects showed that the RF-Waveform classifier performed better than the PLR-Waveform classifier. We did not consider the PLR-Riemann classifier at this stage because its training was too slow to apply between offline and online phases. Therefore, we trained an RF-Waveform classifier for each subject individually on the data obtained in the offline phase and applied it in real time in the online phase. The probability of being a target was estimated on each instance of every symbol along a road segment and sent back to the driving simulator. Based on the probabilities averaged over instances for each of four symbols separately, the BCI system will infer the target symbol at the end of a road segment.

### 3.4. Performance evaluation

#### 3.4.1. Offline performance evaluation

We employ nested cross-validation to adjust various hyperparameters in the inner loop: regularization term for PLR and the tree depth in Random Forest. The purpose of the outer loop is to obtain an unbiased performance estimation, so it is critical to avoid training and testing on correlated data. We achieve it by performing leave-one-run-out for the outer loop, although we had only 3 offline runs. The inner loop is implemented with 4-fold cross-validation while keeping the temporal order of the trials before the split. Since the classes of target and non-target eye fixations are unbalanced, we utilized AUC to measure the classification performance. The group-level average performance was tested with a non-parametric one-sided one-sample Wilcoxon signed-rank test against the theoretical chance level of 0.5 AUC for each classifier. The \(p\)-values are corrected with the Benjamini–Hochberg correction \([42]\). The differences between classifiers are estimated with a non-parametric Friedman test.

#### 3.4.2. Simulated online performance evaluation

After training the classifiers on all the offline data, we applied them to all the online data and obtained a single AUC value. The applied statistical procedure is the same as for the offline phase data.

#### 3.4.3. Online performance evaluation

During the online phase, we predicted the target symbol from the EFRP classification. We assess the overall performance with accuracy and confusion matrices for four symbols. Accuracies were estimated per subject and averaged across subjects. The significance of single-subject performance was tested with the exact binomial test against 0.25 (the theoretical chance level with four balanced classes) \([43]\). Due to the high number of tests, we applied the Benjamini–Hochberg correction \([42]\). The confusion matrices were tested for independence with Pearson’s chi-square test \([44]\) separately for each subject with Benjamini–Hochberg correction.

As for the level of statistical significance, we select the standard level of \(p = 0.05\).

### 4. Results

#### 4.1. Gaze analysis

4.1.1. Comparison of fixation extraction methods

I2MC vs. IDT

We used two different methods to extract fixations in the offline and online phase of the experiment. We compared them using the offline phase data via two measures: (1) the average number of all extracted dwells on boards per run and (2) the average dwell time (table 2). The number of dwells obtained with IDT method is 10% smaller (260 vs. 284), which is a statistically significant difference \((p < 0.0001) with one-tailed t-test\). Since we used the IDT method in the online phase, we missed 10% of the fixations, which led to either missed attendance or labeling some of the second fixations as first fixations. The dwell duration is slightly longer with the IDT method (433 vs. 408 ms), and it is not significantly different from the I2MC method \((p = 0.06 with one-tailed t-test)\).

<table>
<thead>
<tr>
<th></th>
<th># of dwells</th>
<th>Dwell time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I2MC</td>
<td>284 ± 59</td>
<td>408 ± 54</td>
</tr>
<tr>
<td>IDT</td>
<td>260 ± 64</td>
<td>433 ± 86</td>
</tr>
</tbody>
</table>

4.1.2. Board attendance

For the analysis of attendance rate, we used the IDT method to extract fixations for the data from both offline and online phases. In the offline phase, subjects visually attended most of the boards. The attendance rate is almost halved in the online phase. The average attendance rate is shown in table 3. Two-way repeated measures ANOVA shows a minor significant effect in the interaction (experimental phase × board type) with \(F_{(1,11)} = 8.927, p = 0.012\). The main effect of board type (targets vs. non-target) is also minor but significant with \(F_{(1,11)} = 4.923, p = 0.0485\). However, the experimental phase (offline vs. online) shows a major and significant effect with \(F_{(1,11)} = 405, p < \)
4.1.3. Counting
The total number of targets in the offline phase is 173. We analyzed the button presses after each road segment, which should be equal to the number of targets on each road. The average number of incorrect counts (both missed and extra counts) was 5, the worst performance was at 15 errors (see the number of miscounts in figure 7).

4.1.4. Dwell time
We analyzed the dwell time distributions on targets vs. non-targets in the offline and online phases (figure 4). Most of the dwells are limited to the time between the boards pop up equal to 900 ms. The dwell times are identical for non-targets in both phases and statistically significantly shorter than for targets ($p < 0.0001$). The median dwell time for targets is statistically significantly longer in the online phase ($p < 0.0001$).

4.2. EFRP waveform
We present the analysis of the aggregated EFRP waveform for the four subjects (S1–S4 is available online at stacks.iop.org/JNE/18/026010/mmedia) who demonstrated the highest classification performance in the offline phase with the PLR-Waveform classifier because it allows for direct representation of the features used by a linear classifier (see section 4.3, figure 7). The EFRP waveform results for all subjects are provided in the supplementary material (cite here).

We visualized the representative Cz channel signal for each eye fixation while ordering them by the dwell time (figure 5). The amplitude of the presented EFRP is limited to the range $[-1.5, 1.5]$ $\mu$V. The complex of components right after the fixation onset ranging from 100 to 300 ms reflects the evoked potentials from the fixation itself. It contains negative and positive deflections. We can observe the same complex of components after the dwell offset. The shift of gaze happens right where we expect the P300-like component, so it can be masked by this evoked activity. The positive deflection occurs at the end of the dwell for both targets and non-targets, however, it has a greater amplitude for targets as seen on the averaged EFRP.

The spatio-temporal distribution of discriminant power is visualized with signed R-squared ($r^2$) value obtained on aggregated epochs of the four best performing subjects (figure 6). The results are qualitatively similar between the offline and online phases. The greatest values are mainly confined within the region between 100 and 700 ms. There is a spatio-temporal evolution of discriminant activity spreading across the whole scalp. It is expected as the visual processing involves different brain regions. It is important to emphasize that the interpretation of discriminant activity’s spatial distribution depends on the chosen referencing method (CAR in our case). Nonetheless, the P300-like component of EFRP can be seen at 500 ms after the fixation onset.

4.3. Comparison of decoding approaches
4.3.1. Offline
All the classification methods yielded a single-trial performance between 0.52 and 0.60 AUC on average (figure 7), which is statistically significant against the chance level of 0.5 for all methods ($p < 0.05$ with Wilcoxon tests after Benjamini–Hochberg correction for six methods). Five out of 12 subjects achieve performance above 0.6 for at least one of the approaches based on EEG features. The performance of different approaches has a different ranking per subject, and the differences between the approaches are statistically significant ($\chi^2 = 6.5$, $p = 0.0001$ with repeated measures ANOVA). It is worth noticing that the combination of both dwell time and EEG waveform is better than just one of these feature sets with the post hoc one-sided Wilcoxon signed-rank providing $p = 0.04$. The Riemann feature set has poorer performance than the Riemann+ feature set with the post hoc one-sided Wilcoxon signed-rank providing $p = 0.001$.

4.3.2. Simulated online
The average AUC values lie between 0.55 and 0.73 for each approach, which is statistically significant against 0.5 for all approaches. ($p < 0.05$ with Wilcoxon tests after Benjamini–Hochberg correction for six approaches).

It is worth noting that the performance of EEG-based approaches on online data is consistent with the training performance on offline data. However, for approaches relying on the dwell time, the performance drastically improved for several subjects compared to offline performance. The average AUC for the Dwel classifier increased from 57 to 73 ($p = 0.0002$ with one-sided Wilcoxon signed-rank) and for the PLR-Combined classifier from 60 to 67 ($p = 0.008$ with one-sided Wilcoxon signed-rank). This improvement is a direct consequence of the increased target dwell time together while non-target

<table>
<thead>
<tr>
<th></th>
<th>Offline</th>
<th>Online</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targets</td>
<td>$0.92 \pm 0.07$ (173)</td>
<td>$0.47 \pm 0.06$ (213 ± 39)</td>
</tr>
<tr>
<td>Non-Targets</td>
<td>$0.92 \pm 0.07$ (547)</td>
<td>$0.44 \pm 0.09$ (635 ± 120)</td>
</tr>
</tbody>
</table>

0.0001. Since the effect of board type is statistically significant, we can assume that subjects could sometimes differentiate between targets and non-targets without directly looking at the board. However, due to the small difference in the attendance rate, this effect is negligible in the context of our study.

Table 3. The board attendance rate for targets and non-targets in the offline and online phases with IDT method of fixation extraction. The values represent the average and standard deviation across 12 subjects. The values in parentheses represent the number of boards per subject.

J. Neural Eng. 18 (2021) 026010 R. Aydarkhanov
dwell time remained unchanged, see figure 4. However, the underlying question of why the dwell time changed between the offline and the online phases remains.

4.4. Online performance
We assessed each subject’s task performance as the accuracy of detecting the target symbol at the end of a road segment based on the evidence obtained by applying an RF-Waveform classifier to decode driver’s brain responses to multiple instances of symbols along a road segment (table 4). The averaged accuracy equals 0.37, and it is statistically significantly above the chance level of 0.25 for a multi-class classification with four balanced classes ($p < 0.002$ with one-sided one-sample Wilcoxon signed-rank test). Additionally, we applied a statistical test to assess the accuracy per subject. After adjusting $p$-values

![Figure 4. Dwell time distribution for first dwells on targets vs. non-targets in the offline and online phases. The single-subject distributions were smoothed, normalized and averaged across subjects.](image)

![Figure 5. The signal of Cz channel for extracted EFRP epochs aggregated for the four best performing subjects with the offline AUC above 0.6 with the PLR-Waveform approach. Left: offline phase. Right: online phase. Top panel: target EFRP epochs. Middle panel: non-target EFRP epochs. Bottom line plot: average signal across target and non-target EFRP epochs. The epochs are ordered according to the dwell time shown with the black S-shaped curve.](image)
Figure 6. The discriminant power for the aggregated epochs of the four best performing subjects with the offline AUC above 0.6 with the PLR-Waveform approach. Signed $R^2$ is demonstrated for midline channels around the eye fixation onset (top left: offline, top right: online) and on topographic maps (top map: offline, bottom map: online).

Figure 7. Performance of EFRP classification with various approaches in the offline analysis (top) and the simulated online analysis (bottom) as a percentage of AUC. Each dot shows single fold performance in a leave-one-run-out cross-validation for the corresponding classification approach. The overlaid gray bars at the top show the behavioral performance: the number of mis-counts in the offline phase with the respect to the right-hand y-axis.

with the Benjamini–Hochberg procedure, 8 out of 12 subjects performed statistically above chance level.

To verify the independence of the four classes in the online phase, we computed the aggregated confusion matrices across all symbols for all subjects (figure 8). We performed an independence test for confusion matrices per subject. After adjusting $p$-values with the Benjamini–Hochberg procedure, only two subjects show a significant imbalance in the confusion matrix. For these two subjects, the accuracy of the symbol $\Xi$ was higher, and the accuracy of the symbol $E$ was lower than of the other symbols.

5. Discussion

The integration of the BCI systems in the daily life of healthy/able-bodied users requires the system to be built around the experimental paradigms supporting natural human behavior. To this end, EFRP-based decoding of cognitive processes in overt visual search has the potential to augment human-machine interaction. In this study, we investigated the decoding of visual recognition in a driving scenario. It resembles one of the typical everyday activities and provides the associated challenges in decoding visual recognition: free eye gazes, dynamic visual input,
primary tasks. For this purpose, we limit the driving task to following the simple route at a comfortable and natural speed. To avoid overloading subjects’ attention with too many distractors, we did not include other participants on the road nor moving objects. Nonetheless, due to the car’s movement, the drivers were subjected to a dynamic visual input and perceived a natural optic flow.

5.1. Ocular behavior in driving
We considered the random pop-up appearance of the task-relevant stimuli above the sidewalks. In the previous study on active search in a dynamic scene [29], this type of appearance created sufficient locking of the cognitive EFRP’s components to the fixation onset. At the same time, their results showed that the time spent on the stimuli was not informative about the stimuli type (target vs. non-target) in contrast to the more attentionally demanding motion appearance conditions. Interestingly, we observed the time spent on the stimuli to be discriminative in our experimental scenario, although the same pop-up stimuli appearance was considered. It can be explained by the effect of the user’s/driver’s motion relative to the objects on the visual information processing and the attentional load. Moreover, in the online phase, subjects looked longer on targets compared to the offline phase. Although the subjects’ dwell time was not decoded directly, they were aware that they could potentially influence the decoding quality. It might lead to deliberate or unconscious changes in their behavior. Some subjects could achieve a high decoding performance based only on dwell time in the offline phase. But with the changes in the behavior during the online phase, most of the subjects drastically improve in their decoding performance based on the dwell time.

The board attendance rate reveals two aspects of the task completion: (1) how well the subjects coped with the pace and attentional load and (2) whether they can differentiate between targets and non-targets using their peripheral vision. We observed a statistically significant difference in the board attendance between offline and online phases despite the equal number of boards. On the one hand, we expected the subjects to be more engaged in the task due to the interactive feedback part. However, the observed increase of the dwell time on target boards in the online phase made it more challenging to attend all the boards within the limited time. On the other hand, behavioral response based on the number of recognized targets was not required, which might relax the cognitive load.

We carefully designed stimuli to ensure that their recognition requires foveal vision. It is confirmed by the similar attendance rate on targets and non-targets. However, the difference is statistically significant, suggesting that peripheral recognition is not completely excluded with the provided stimuli design. We were concerned that introducing new symbols in the online phase may change the behavior or the cognitive response due to the novelty. The balanced confusion matrices of decoding target recognition in the online phase confirm that subjects perceived all the symbols equally with regard to the task. However, the presence of additional types of stimuli may have contributed to the increased dwell time on the target boards because the target identification among four symbols closely resembling each other is more challenging than among only two symbols [45, 46].
Table 4. Summary of online performance: accuracy of target decoding per road segment and p-values for two statistical tests. The accuracy is tested with binomial distribution. The recognition of a target symbol independent of the actual symbol is tested with a chi-square test of independence. The p-values are corrected with Benjamini–Hochberg method per row. Statistically significant results are typed in bold.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
<th>S10</th>
<th>S11</th>
<th>S12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.44</td>
<td>0.38</td>
<td>0.45</td>
<td>0.29</td>
<td>0.44</td>
<td>0.35</td>
<td>0.39</td>
<td>0.25</td>
<td>0.31</td>
<td>0.38</td>
<td>0.38</td>
<td>0.4</td>
</tr>
<tr>
<td>Accuracy test</td>
<td>0.001</td>
<td>0.04</td>
<td>0.001</td>
<td>0.48</td>
<td>0.001</td>
<td>0.07</td>
<td>0.02</td>
<td>1.0</td>
<td>0.24</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Independence test</td>
<td>0.004</td>
<td>0.29</td>
<td>0.03</td>
<td>0.74</td>
<td>0.08</td>
<td>0.17</td>
<td>0.1</td>
<td>0.41</td>
<td>0.48</td>
<td>0.19</td>
<td>0.22</td>
<td>0.26</td>
</tr>
</tbody>
</table>

5.2. Decoding visual recognition in driving

The discriminant analysis of EFRP shows similar results for both offline and online phases. Most of the relevant features lie within the [200, 700] ms window, which coincides with the dwell durations. The spatial localization of relevant features is consistent with the typical spatial distribution of the P300 component in the oddball paradigm. The EFRP waveforms are known to contain a strong P1 component at the occipital area that reflects the beginning of the visual processing of a stable visual input after the saccade. In the analysis of the Cz channel, it corresponds to the negative deflection at 100 ms. It is clearly present in most of the first fixations on boards. Moreover, we could also see it for the following fixations. During the dwell time, multiple fixations can be detected, leading to the overlap of the P300-like component locked to the first fixation and the following early EFRP component. In contrast to the classical P300 paradigms where the the visual input timing is controlled (with a constant frequency), here, the overlap of EFRP is variable across trials. It is an additional challenge as compared to typical ERP detection approaches. Removing the activity related to previous and subsequent fixations from the EEG was attempted by modeling it from various characteristics of the previous and subsequent fixations. [28, 30, 47]. We did not choose to apply this approach as there is a risk of distorting the signal of interest. The precise EFRP shape depends on preceding and following eye behavior (e.g. amplitude and direction of the saccades) as well as low-level features of the visual input [48].

We compared multiple EFRP-based classification approaches on the offline data. In addition, we did a comparative study using recorded online data (simulated online analysis). All approaches, including waveform-based linear and non-linear and covariance-based methods, resulted in a similar performance on average across subjects, which is significantly above the chance level for all the approaches. The combination of waveform features with the dwell time outperforms other feature sets on average. However, there is no single best approach for all subjects.

The actual online closed-loop performance measured by the target symbol identification (a four-class problem) is significantly above the chance level for eight subjects. Moreover, on average across subjects, the accuracy of 0.37 is statistically significantly higher than the chance level.

Considering the observed dwell-time difference between target and non-target boards, one can argue that the relatively good performance of EFRP classifiers is due to the early evoked components of the following fixations rather than due to the later cognitive components of the analyzed EFRP. Therefore, these classifiers may indirectly rely on the difference between the target and non-targets dwell times. However, this was not the case as the simulated online performance of the EEG-based decoders did not exhibit an improvement similar to the classification models based on dwell time.

A critical part of real-time systems such as the one implemented in our experiments is the computational cost of the used algorithms. The approaches used in the online phase proved to be time-efficient without causing delays. Eye fixation extraction was successfully done with IDT provided by SMI. I2MC method requires more computation due to a clustering step on a sliding window manner, however, it might also be adjusted for a real-time application. Spatial and temporal filtering, as well as eye movement artifact removal, are fast linear operations. Regarding feature extraction and inference of the classifier, these steps are the least demanding in computational time because they are applied only as frequently as the relevant eye fixations are detected. All mentioned pairs of features and classifiers were applicable in the current setup, therefore, it did not affect the choice of the approach in the online phase. Nonetheless, the training even on a single subject data required a considerable amount of time, especially when applied multiple times within nested cross-validation. The most demanding steps included the estimation of parameters for the Riemannian features and the Random Forest classifier. This led to an extended pause between the offline and online phases of the experimental session.

One of the challenges in building a real-world BCI application is an adequate training protocol design. The training protocol has to secure good quality data for training a classifier. Thus, the practice is to create highly controlled trials in terms of stimuli and the subjects’ behavior. At the same time, the introduced constraints should not compromise the data not to be representative of the actual phenomenon in real-world conditions when operating the BCI application. It is known that stimuli—particularly their semantic richness, ambiguity, and diversity—may
affect visual recognition and the corresponding ERP components. We believe that the approach to address this challenge should include two actions: (i) to create a training protocol as close as possible to the real-world setup, and (ii) to consider decoding algorithms able to deal with variability in neural responses [49]. We plan to further investigate in this direction and encourage the other researchers in this field to address this challenge.

Finally, in this study, our focus was on detecting the EEG correlates of object recognition in free viewing visual search tasks during driving. Ocular correlates such as pupil dilation along with multimodal integration of neural and ocular correlates may be further investigated in future studies. Multimodal integration may improve decoding performance [15].

6. Conclusion

In this study, we have demonstrated the feasibility of decoding visual target recognition from brain signals in challenging conditions of a naturalistic driving scenario. By decoding the brain signals, the smart car can acquire information about the objects of drivers’ interest and enhance their interaction by creating a tailored recommendation. Considering that no extra effort is required from the driver and that he could quickly accept or ignore the recommendations, the extremely high decoding performance is not a prerequisite. Our results are, therefore, promising for bringing this type of BCIs to smart cars. Finally, our approach can benefit from improvements in different components, such as decoders robust to the natural temporal variability of EEG potentials, reliable and computationally efficient eye fixation algorithms, longer training of subjects over multiple sessions, and multimodal integration of neural and ocular data. Bigger datasets would also allow the application of more powerful machine learning models, e.g. deep neural networks.

Acknowledgment

The authors thank Nissan Motor Co. Ltd for supporting this work.

ORCID iDs

Ruslan Aydarkhanov https://orcid.org/0000-0001-6888-4891
Ricardo Chavarriaga https://orcid.org/0000-0002-8879-2860
José del R Millán https://orcid.org/0000-0001-5819-1522

References


[28] Devillers Hène, Guyader N and Anne G-Dé 2015 An eye fixation–related potentials analysis of the P300 potential for fixations onto a target object when exploring natural scenes J. Vis. 15 20


[32] Kaminenkowksi J E, Ison M J, Quiroga R Q and Sigman M 2012 Fixation-related potentials in visual search: a combined EEG and eye tracking study J. Vis. 12 1


[34] Pena A M, Nirhbar E H, Pacholla A, Chaspari T and Ragan E D 2019 Detecting changes in user behavior to understand interaction provenance during visual data analysis IUI Workshops


[38] Demirali T, Ademoglu A, Comerchero M and Polich J 2001 Wavelet analysis of P3a and P3b Brain Topography 13 251–67


[44] Karl Pearson X 1900 On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling London, Edinburgh Dublin Phil. Mag. J. Sci. 50 157–75

[45] Dick A O 1971 Processing time for naming and categorization of letters and numbers Perception Psychophys. 9 350–2


[47] Kristensen E, Guerin-Dugué A and Rivet B 2015 Comparison between adjar and xDawn algorithms to estimate eye-fixation related potentials distorted by overlapping 2015 7th Int. IEEE/Conf. on Neural Engineering (NER) pp 976–9


Mobile brain/body imaging of landmark-based navigation with high-density EEG

Alexandre Delaux1 | Jean-Baptiste de Saint Aubert1 | Stephen Ramanouël1 | Marcia Bécu1 | Lukas Gehrke2 | Marius Klug2 | Ricardo Chavarriaga3,4 | José-Alain Sahel1,5,6,7 | Klaus Gramann2 | Angelo Arleo1

Abstract

Coupling behavioral measures and brain imaging in naturalistic, ecological conditions is key to comprehend the neural bases of spatial navigation. This highly integrative function encompasses sensorimotor, cognitive, and executive processes that jointly mediate active exploration and spatial learning. However, most neuroimaging approaches in humans are based on static, motion-constrained paradigms and they do not account for all these processes, in particular multisensory integration. Following the Mobile Brain/Body Imaging approach, we aimed to explore the cortical correlates of landmark-based navigation in actively behaving young adults, solving a Y-maze task in immersive virtual reality. EEG analysis identified a set of brain areas matching state-of-the-art brain imaging literature of landmark-based navigation. Spatial behavior in mobile conditions additionally involved sensorimotor areas related to motor execution and proprioception usually overlooked in static fMRI paradigms. Expectedly, we located a cortical source in or near the posterior cingulate, in line with the engagement of the retrosplenial complex in spatial reorientation. Consistent with its role in visuo-spatial processing and coding, we observed an alpha-power desynchronization while participants gathered visual information. We also hypothesized behavior-dependent modulations of the cortical signal during navigation. Despite finding few differences between the encoding and retrieval phases of the task, we identified transient time–frequency patterns attributed, for instance, to attentional demand, as reflected in the alpha/gamma range, or memory...
workload in the delta/theta range. We confirmed that combining mobile high-density EEG and biometric measures can help unravel the brain structures and the neural modulations subtending ecological landmark-based navigation.

**KEYWORDS**
ecological navigation, mobile EEG, retrosplenial complex, source reconstruction, virtual reality

1 | INTRODUCTION

Spatial navigation requires active exploration, multisensory integration, as well as the encoding and long-term consolidation of internal models of the world (Arleó & Rondi-Reig, 2007; Epstein et al., 2017; Wolbers & Hegarty, 2010). Thus, the ability to navigate in space encompasses both perceptual and cognitive faculties (Ekstrom et al., 2017; Spiers & Barry, 2015). A large body of work has elucidated the neural bases of wayfinding behavior in both animals and humans, leading to a better understanding of the navigational system across multiple levels (Burgess, 2008; Epstein et al., 2017; Hardcastle et al., 2017; Poulter et al., 2018).

Most investigations of the brain network subverting human spatial navigation rely on functional magnetic resonance imaging (fMRI) (Epstein et al., 2017; Taube et al., 2013) due to its unmatched spatial resolution among non-invasive methods. However, this technique is not suited for testing participants in unconstrained motion conditions, which limits the study of neural processes involved during natural behavior (Zaitsev et al., 2015). Combining behaviormetric and neuroimaging recordings in ecological (i.e., close to real, natural) conditions is key to modern cognitive neuroscience (Ladouce et al., 2019; Schaefer, 2014), in particular to study spatial cognition (Bécu et al., 2020a; Gehrke & Gramann, 2021; Miyakoshi et al., 2021). With relatively coarse spatial but fine temporal resolution, electroencephalography (EEG) offers a complementary tool for neuroimaging the brain during spatial behavior (Baker & Holroyd, 2009; Bischof & Boulanger, 2003; Lin et al., 2009, 2015; Plank et al., 2010). Although EEG does not prevent the participant's motion per se, it is very sensitive to movement-related artifacts. Electrical potentials from muscle contractions (e.g., head movements, eye blinks, or heartbeat, see Jung et al., 2000) generate strong artifactual signals that compromise the extraction of brain-related responses (i.e., reducing the signal-to-noise ratio). As a consequence, most EEG studies have constrained the mobility of participants in order to minimize motion-related artifacts (e.g., by making them sit in front of a screen and respond with finger taps only).

Recent technical developments have unlocked the possibility of using EEG brain imaging in a variety of ecological conditions (indoor walking: Luu et al., 2017a; Ladouce et al., 2019; Park & Donaldson, 2019; outdoor walking: Debenere et al., 2012; Reiser et al., 2019; cycling: Zink et al., 2016; di Fronso et al., 2019; and dual tasking: Marcar et al., 2014; Bohle et al., 2019). By coupling EEG recordings with other biometric measures (e.g., body and eye movements), the Mobile Brain/Body Imaging (MoBI) paradigm has been successfully combined with fully immersive virtual reality (VR) protocols (Djebbara et al., 2019; Liang et al., 2018; Peterson & Ferris, 2019; Plank et al., 2015; Snider et al., 2013). Immersive VR allows near-naturalistic conditions to be reproduced, while controlling all environmental parameters (Diersch & Wolbers, 2019; Park et al., 2018; Parsons, 2015; Starrett & Ekstrom, 2018). The reliability of 3D-immersive VR enables the stimulation of visual, auditory, and proprioceptive modalities, while allowing the participant to actively explore and sense the virtual environment (Bohil et al., 2011; Kober et al., 2012). This continuous interplay between locomotion and multisensory perception is thought to be a key component of spatial cognition in near-natural conditions, as its absence leads to impaired performance in various spatial abilities (path integration: Chance et al., 1998; spatial updating: Klier & Angelaki, 2008; spatial reference frame computation: Gramann, 2013; spatial navigation and orientation: Taube et al., 2013; Ladouce et al., 2017; and spatial memory: Holmes et al., 2018).

In the present study, we use the MoBI approach to combine high-density mobile EEG recordings and immersive VR in order to study spatial navigation in a three-arm maze (i.e., a Y-maze). Our primary aim is to provide a proof-of-concept in terms of EEG-grounded neural substrates of landmark-based navigation consistent with those found in similar fMRI paradigms (Iaria et al., 2003; Konishi et al., 2013; Wolbers & Büchel, 2005; Wolbers et al., 2004). We chose the Y-maze task because it offers a simple two-choice behavioral paradigm suitable to study landmark-based spatial navigation and to discriminate between allocentric (i.e., world-centered) and egocentric (i.e., self-centered) responses, as previously shown in animals (Barnes et al., 1980) and humans (Bécu et al., 2020b; Rodgers et al., 2012). Complementarily, a recent fMRI study of ours has investigated the brain activity of regions involved in visuo-spatial processing and navigation in a similar Y-maze task (Ramanöel et al., 2020). This offers the opportunity to comparatively validate the neural correlates emerged through static fMRI experiment against those found by mobile high-density EEG.
The neural substrates of landmark-based navigation form a network spanning medial temporal areas (e.g., hippocampus and para-hippocampal cortex) and medial parietal regions (Epstein & Vass, 2014), such as the functionally defined retrosplenial complex (RSC) (Epstein, 2008). Here, we expect the RSC to play a role in mediating spatial orientation through the encoding and retrieval of visual landmarks (Auger & Maguire, 2018; Auger et al., 2012, 2015; Julian et al., 2018; Marchette et al., 2015; Spiers & Maguire, 2006). The RSC is indeed implicated in the translation between landmark-based representations in both egocentric and allocentric reference frames (Marchette et al., 2014; Mitchell et al., 2018; Shine et al., 2016; Sulpizio et al., 2013; Vann et al., 2009). Our hypothesis also encompasses the role of specific upstream, visual processing areas of the parieto-occipital region involved in active wayfinding behavior (Bonner & Epstein, 2017; Patai & Spiers, 2017). In addition, our paradigm accounts for the role of downstream, higher-order cognitive functions necessary for path evaluation, covering a frontoparietal network (including prefrontal areas, Epstein et al., 2017; Spiers & Gilbert, 2015) that codes for overarching mechanisms such as spatial attention and spatial working memory (Cona & Scarpazza, 2019).

Mobile brain imaging protocols also engage locomotion control processes, in which motor areas in the frontal lobe and somatosensory areas in the parietal lobe are typically involved (Gwin et al., 2010; Seeber et al., 2014; Roeder et al., 2018; see Delval et al., 2020 for a recent review). Furthermore, the integration of vestibular and proprioceptive cues made possible by mobile EEG paradigms is likely to influence the observed neural correlates of spatial orientation (Ehinger et al., 2014; Gramann et al., 2018) and attention (Ladouce et al., 2019).

Finally, given the high temporal resolution of EEG, we aim at characterizing how the activity of the structures engaged in active, multimodal landmark-based navigation is modulated by behavioral events, related to either action planning (e.g., observation of the environment, physical rotation to complement mental perspective taking) or action execution (e.g., walking, maintaining balance). We also aim at exploring the differential engagement of brain regions involved in the encoding (learning condition) and the retrieval (control and probe conditions) phases of the task (RSC is implicated in both; Burles et al., 2018; Epstein & Vass, 2014; Mitchell et al., 2018).

The purpose of this study is thus to explore the cortical correlates of landmark-based navigation in mobile participants. We first hypothesize that the analysis of the EEG signal will retrieve the above-mentioned brain structures known to be engaged during active spatial navigation based on visual cues. We then expect behavioral events to modulate features of the recorded EEG data, identifiable as transient time–frequency patterns in the involved brain areas, and to interpret them with respect to spatial cognition and locomotion control literature. Finally, we expect to find significant differences in these patterns across the phases of the task, contrasting the cognitive mechanisms involved in context-dependent task solving. We aim to investigate and interpret their condition specificity and their temporality. Under such considerations, this work can help toward a better understanding of context-specific neural signatures of landmark-based navigation.

2 METHODS

2.1 Participants

Seventeen healthy adults (range: 21–35 years old, \(M = 26.82, SD = 4.85; 10 \text{ women} \)) participated to this study. Fifteen were right-handed and two left-handed. All participants had normal (or corrected to normal) vision and no history of neurological disease. In one recording session, there were abnormalities (discontinuities and absence of events) in the motion capture signal. Thus, we removed one participant from the analysis. The experimental procedures were approved by the local ethics committee (GR_12_20190513, Institute of Psychology & Ergonomics, Technische Universität Berlin, Germany) and all participants signed a written informed consent, in accordance with the Declaration of Helsinki. All participants answered a discomfort questionnaire at the end of the experiment, adapted from the simulation sickness questionnaire of Kennedy et al. (1993), which can be found in Methods S1. We gave the instructions in English and all participants reported a good understanding of the English language. Each participant received a compensation of either 10€/h or course credits.

2.2 EEG system

The EEG system (Figure 1a) consisted of 128 active wet electrodes (actiCAP slim, Brain Products, Gilching, Germany) mounted on an elastic cap with an equidistant layout (EASYCAP, Herrsching, Germany). The impedance of a majority of the channels was below 25 kΩ (9.5% of the electrodes had an impedance above 25 kΩ). Two electrodes placed below the participant’s eyes recorded electro-oculograms (EOG). An additional electrode located closest to the standard position F3 (10–20 international system) provided the reference for all other electrodes. The EEG recordings occurred at a sampling rate of 1 kHz. The raw EEG signal was streamed wirelessly (BrainAmp Move System, Brain Products, Gilching, Germany) and it was recorded continuously for the entire duration of the experiment.
FIGURE 1  Virtual environment, setup, and timeline of the experiment. (a) Details of participant’s equipment. (1) EEG cap (128 channels); (2) VR Head-mounted display (VIVE Pro); (3) Wifi transmitter for EEG data (Move system); (4) Additional motion capture tracker (VIVE tracker); and (5) Backpack computer running the virtual environment (Zotac PC). (b) Virtual environment. Participants explored a virtual equilateral Y-maze. In the learning condition, they always started in the same arm (e.g., A) and they had to find a hidden goal, always placed in the same location (e.g., C). In the testing conditions, the environment and goal location stayed the same but the participant would start from either the same position (A) in control trials or the third arm (B) in probe trials. (c) Spatial discretization of the environment (example for a learning trial). We delimited 10 areas in the maze: “S” stands for starting arm, “C” for center, “E” for error arm, and “G” for goal arm. In the text, when referring to the arm chosen by the participant (either “E” or “G”), we use the letter “F” standing for finish arm. These labels are condition-dependent (different in the probe condition). The names of the landmark depend on the location of starting arm in the learning condition and goal arm. These names are block dependent. (d) General timeline of the experiment. The first row represents the general succession of conditions in the experiment. The second row shows an example of the sequence of trials in an experimental block. The third row illustrates the structure of a trial, including a possible course of events: progress across spatial sections and visibility of landmarks depending on participant’s head movements. We provide a video of a participant performing the task, along with the reconstruction of the tracker positions, in Video S1.
2.3 Virtual Y-maze and motion tracking

The virtual maze consisted of an equilateral Y-maze (3-armed maze) with three distal landmarks placed outside the maze, 20 m away from the center and visible above the walls (Figure 1b). The landmarks were abstract geometric shapes (e.g., square, circle, star). The wall texture and the light were homogeneous and non-informative. Each arm of the maze was 90 cm wide and 225 cm long. For the sake of analysis, the maze was discretized into 10 zones (3 evenly divided zones per arm and one for the maze center, Figure 1c). These zones were not visible to the participant. Crossing between zones was recorded online without influencing the task flow.

We designed the virtual Y-maze by using the Unity3D game engine (Unity Technologies, San Francisco, California, USA, version 2017.1.1f1 for Windows), and we rendered it using an HTC Vive Pro head-mounted display (HTC Corporation, Taoyuan, Taiwan) with a 90 Hz refresh rate (2 times AMOLED 3.5” 1440x1600 pixels, 615 ppi, and 110° nominal field of view). The HTC was connected to a VR capable backpack computer (Zotac PC, Intel 7th Gen Kaby Lake processor, GeForce GTX 1060 graphics, 32GB DDR4-2400 memory support, Windows 10 OS, ZOTAC Technology Limited, Fo Tan, Hong Kong) running on batteries and controlled remotely (Figure 1a). An integrated HTC Lighthouse motion tracking system (four cameras, 90 Hz sampling rate, covering an 8 x 12 m area) enabled the recording of the participant’s head by tracking the HTC Vive Pro head-mounted display. It also enabled the tracking of the torso movements via an additional HTC Vive Track placed on the participant’s backpack. We virtually translated the position of this tracker to better reflect the real position of the participant’s torso by considering his or her body measurements. The torso tracker was also used to trigger spatial events (e.g., reaching the goal, crossing spatial section boundaries). The height of the maze walls and the altitude of landmarks were adjusted to the participant’s height (based on the head tracker) to provide each participant with the same visual experience. Each participant wore earphones playing a continuous white noise to avoid auditory cues from the external world. During the disorientation periods (see protocol), relaxing music replaced the white noise. One experimenter gave instructions through the earphones, while monitoring the experiment from a control room. The participant could answer through an integrated microphone. He/she was instructed to refrain from talking while performing the experiment to limit artifacts in the recorded EEG signal. Another experimenter stayed with the participant inside the experimental room to help with potential technical issues and conduct the disorientation, avoiding any interaction with participants during the task. The EEG signal, motion capture, and all trigger events were recorded and synchronized using the Lab Streaming Layer software (Kothe, 2014).

2.4 Experimental protocol

An entire experimental session lasted 3 hr on average and it included preparing the participant with the EEG and VR equipment and running the experimental protocol. The immersion time in VR was between 60 and 90 min.

2.4.1 Free exploration phase

Before starting the actual task, the participant explored the Y-maze for 3 min, starting at the center of the maze. He/she was instructed to inspect all details of the environment and to keep walking until the time elapsed. The purpose of this phase was to familiarize the participant to the VR system and the Y-maze environment (including the constellation of landmarks).

2.4.2 Navigation task

The navigation task included a learning condition and a testing condition. During learning, the participant began each trial from the starting arm (e.g., location A in Figure 1b) and he/she had to find the direct route to a hidden target at the end of the goal arm (e.g., location C in Figure 1b). Upon reaching the goal, a reward materialized in front of the participant (3D object on a small pillar representing, for instance, a treasure chest) to indicate the correct location and the end of the current trial. The learning period lasted until the participant reached the goal directly, without entering the other arm, three times in a row. Before each trial, we disoriented the participant to ensure that he/she would not rely on previous trials or the physical world to retrieve his/her position and orientation. To disorient the participant, the experimenter simply walked him/her around for a few seconds with both eyes closed (and the head-mounted display showing a black screen). The testing condition included six trials: three control trials and three probe trials, ordered pseudo-randomly (always starting with a control, but never with three control trials in a row). In the control trials, the participant started from the same arm as in the learning condition (e.g., location A in Figure 1b). In the probe trials, he/she started from the third arm (e.g., location B in Figure 1b). Before starting a new trial (either control or probe), the participant was always disoriented. Then, he/she had to navigate to the arm where he/she expected to find the goal and stop there (without receiving any reward signal). If the participant went to the incorrect arm, it was considered as an error. We present a single trial example of one participant...
performing the task and we illustrate the motion tracking in the virtual environment in Video S1.

### 2.4.3 | Block repetitions

The sequence “learning condition + testing condition” formed an experimental block. Each participant performed nine experimental blocks (Figure 1d). In order to foster a feeling of novelty across block repetitions, we varied several environmental properties at the beginning of each block: wall texture (e.g., brick, wood, etc.), goal location (i.e., in the right or left arm, relative to the starting arm), reward type (e.g., treasure chest, presents, etc.), as well as the shape (e.g., circle, square, triangle, etc.) and color of landmarks. When changing the environment between blocks, we kept the maze layout and landmark locations identical. The sequence of blocks was identical for all participants, who had to take a compulsory break after the fourth block (Figure 1d). In addition, after the sixth or seventh block, a break was introduced when requested by the participant.

### 2.4.4 | EEG baseline recordings

Both before the free exploration period and after the 9th block, the participant had to stand for 3 min with his/her eyes opened in a dark environment. This served to constitute a general baseline for brain activity. Similarly, we recorded the EEG baseline signal (in the dark for a random duration of 2–4 s) before each trial (Figure 1d, bottom). Besides providing a baseline EEG activity specific to each trial, this also allowed the starting trial time (i.e., the appearance time of the maze) to be randomized, thus avoiding any anticipation by the participant.

### 2.5 | Behavioral analysis

All analyses were done with MATLAB (R2017a and R2019a; The MathWorks Inc., Natick, MA, USA), using custom scripts based on the EEGLAB toolbox version 14.1.0b (Delorme & Makeig, 2004), the MoBILAB (Ojeda et al., 2014) open source toolbox, and the BeMoBIL pipeline (Klug et al., 2018).

#### 2.5.1 | Motion capture processing

A set of MoBILAB’s adapted functions enabled the preprocessing of motion capture data. The rigid body measurements from each tracker consisted of (x, y, z) triplets for the position and quaternion quadruplets for the orientation. After the application of a 6 Hz zero-lag low-pass finite impulse response filter, we computed the first time derivative for position of the torso tracker for walking speed extraction and we transformed the orientation data into axis/angle representations. An EEGLAB dataset allowed all preprocessed, synchronized data to be collected, and split into different streams (EEG, Motion Capture) to facilitate EEG-specific analysis based on motion markers.

#### 2.5.2 | Allocentric and egocentric groups

Probe trials served to distinguish between allocentric and egocentric responses by making the participants start from a different arm than the one used in the learning period. We assigned a participant to the allocentric group if he/she reached the goal location in the majority of probe trials (i.e., presumably, by using the landmark array to self-localize and plan his/her trajectory). Conversely, we assigned a participant to the egocentric group when he/she reached the error arm in the majority of probe trials (i.e., by merely repeating the right- or left-turn as memorized during the learning period).

#### 2.5.3 | Time to goal

We assessed the efficacy of the navigation behavior by measuring the “time to goal”, defined as the time required for the participant to finish a trial (equivalent to the “escape latency” in a Morris Water Maze). In learning trials, it corresponded to the time to reach the goal zone and trigger the reward. In test trials, it corresponded to the time to reach the believed goal location in the chosen arm (i.e., entering the G1 or E1 zone in Figure 1c).

#### 2.5.4 | Horizontal head rotations (relative heading)

The participant’s heading was taken as the angle formed by his/her head orientation in the horizontal plane with respect to its torso orientation, aligned with the participant’s sagittal plane. After extracting the head and torso forward vectors from each tracker, computing the signed angle between those vectors’ projections in the horizontal plane provided the heading value.

#### 2.5.5 | Walking speed

The forward velocity component of the torso tracker provided the participant’s walking information. For each trial,
we computed the mean and $SD$ of the forward velocity, and their average for each participant. To evaluate movement onsets and offsets, we compared motion data recorded during the trials against those recorded during the short baseline period before each trial, considered as a reliable resting state for movements. Movement transitions (onsets and offsets) were based on a participant-specific threshold, equal to the resting-state mean plus 3 times the resting-state $SD$. The excluded movement periods were those lasting less than 250 ms and during which motion did not reach another participant-specific threshold, equal to the resting-state mean plus 5 times the resting-state $SD$.

2.5.6 | Landmark visibility

For analysis purposes, we named the three landmarks as Landmark 1, Landmark 2, and Landmark 3 (Figure 1c) and we tracked their visibility within the displayed scene. The participant had a horizontal field of view of 110° and a vertical field of view of 60°. Whenever a landmark appeared in the viewing frustum1 (i.e., in the region of virtual space displayed on the screen) and it was not occluded by any wall, it was considered as visible by the participant. Given the restrained horizontal field of view and the configuration of the VR environment, perceiving more than one landmark at the same time was unlikely.

2.5.7 | Zone-based behavioral analysis

The maze discretization (Figure 1c) provided a coherent basis for analyses across trials and participants. To ensure consistency in the comparison between trials, we selected those trials where the participant followed a straightforward pattern (zone-crossing sequence: S1 → S2 → S3 → C → G3 → G2 → G1). We thus discarded all trials in which the participant went backward while navigating (e.g., during learning, when his/her first choice was toward the error arm, and he/she had to come back to the center in order to go toward the goal arm). To further ensure homogeneity, we also excluded those trials in which the time to goal was unusually long (i.e., by computing the outliers of the time to goal distribution across all participants). These selection criteria kept 1,289 (of a total of 1,394) trials for analysis (see Table S1 for details about the distribution of trials across participants).

Finally, we computed offline an additional event corresponding to the first walking onset of the participant in S1 (see Walking speed paragraph above for movement detection) which was inserted in the delimiting sequence of zone-crossing events. For the sake of simplicity, we used the notations “staticS1” for the period preceding this event, and “mobileS1” for the one that follows, before the participant enters S2. Hence, the complete sequence for each trial was, e.g., staticS1 → mobileS1 → S2 → S3 → C → G3 → G2 → G1.

2.5.8 | Motion capture statistics

The above zone-based discretization framed the analysis of the motion capture metrics mentioned above: walking speed, standard deviation of horizontal head rotations, and landmark visibility. For each trial, we first averaged the value of each motion variable over the period between two events of the zone sequence. Then, for each participant, we averaged these values across trials of the same condition.

To better characterize the participants’ behavior in the maze, we investigated how these metrics would depend on the condition, the spatial zone, the landmark (for landmark visibility only), and the different combinations of those factors. Concerning walking speed and standard deviation of horizontal head rotations, we tested the hypothesis that participants would walk slower and make larger head movements in specific zones of the maze related to the challenge posed by the experimental condition (e.g., taking information in S1 during Learning and stopping in C to look at the constellation during Probe). Concerning landmark visibility, we tested the hypothesis that participants would make a differential use of the three landmarks (i.e., preference for one or two) and that attendance to a landmark would depend on the condition and the location of the participant in the maze (e.g., realignment with a preferred landmark at the center specific to Probe condition).

We used fixed model between factors analyses of variance (ANOVA; balanced design) to assess differences and interactions between conditions, zones, and landmarks in those dependent variables. Specifically, for the landmark visibility, we used a three-way ANOVA with the factors: condition (Learning, Control, Probe), landmark (Landmark 1, Landmark 2, Landmark 3), and zone (e.g., staticS1, mobileS1, S2, S3, Center, F3, F2). Note that “F,” standing here for “finish” arm, can be either G for “goal” or E for “error” as used in Figure 1c, depending on the trial outcome. For the walking speed and the standard deviation of horizontal head rotations, we used a two-way ANOVA with the factors condition and zone. The alpha level for significance was set at 0.01 (more conservative level taking into account that we

---

1In 3D virtual reality and computer graphics, the viewing frustum is defined as the region of virtual space displayed on the screen, and it is a coarse imitation of the “cone of vision” in natural viewing. It takes the form of a truncated rectangular pyramid, defined by the horizontal and vertical field of view and by near and far bounds.
FIGURE 2  Flowchart of the EEG processing pipeline. We first preprocessed EEG data at the individual level (in blue) and, in particular, decomposed the channel data into independent components (ICs) with an adaptive mixture ICA (AMICA) algorithm. We then selected 70 ICs per participant for the clustering procedure (in orange). Finally, we labeled and selected the clusters of interest for an ERSP analysis per condition (in brown). The “Cluster selection” process is described in the “EEG cluster analysis” section of the Results.
are computing three simultaneous ANOVAs on the same dataset). When a significant main effect or interaction was found, we used pairwise t-tests (with Tukey's honest significant difference criterion method for multiple comparison correction) to unravel individual differences between factor or interaction terms.

2.6 | EEG data analysis overview

Figure 2 shows the outline of the data preprocessing and analysis steps.

2.7 | Individual EEG analysis

2.7.1 | Processing

We used the BeMoBIL pipeline to preprocess and clean the EEG data (Klug et al., 2018). This pipeline is fully automated and designed to improve signal-to-noise ratio (SNR) in large-scale mobile EEG datasets, which ensures full replicability of the procedure. We first downsampled the data to 250 Hz, applied a 1 Hz high-pass filter to suppress slow drifts in EEG data (zero-phase Hamming windowed finite impulse response filter with 0.5 Hz cut-off frequency and 1 Hz transition bandwidth), and removed spectral peaks at 50 Hz and 90 Hz, corresponding to power line frequency and VIVE headset refreshing rate, respectively (implemented by the cleanLi- neNoise function from the PREP pipeline, Bigdely-Shamlo et al., 2015). We identified noisy channels with automated rejection functions, setting parameters numerical values according to default recommendations from Bigdely-Shamlo et al. (2015). We then reconstructed the removed channels by spherical interpolation of neighboring channels and applied re-referencing to the common average. In a subsequent time-domain cleaning, we detected and removed segments with noisy data. We present more details on the implementation of the cleaning steps in Methods S2.

On the cleaned dataset, we performed an independent component analysis (ICA) using an adaptive mixture independent component analysis (AMICA) algorithm (Palmer et al., 2008), preceded by a principal component analysis reduction to the remaining rank of the dataset taking into account the number of channels interpolated and the re-referencing to the common average. For each independent component (IC), we computed an equivalent current dipole model (ECD) with the DIPFIT plugin for EEGLAB (version 3.0) (Oostenveld & Oostendorp, 2002). For this purpose, we used a common electrode location file obtained from the average of previous measures on participants wearing the same cap. We co-registered this file with a boundary element head model based on the MNI brain (Montreal Neurological Institute, MNI, Montreal, QC, Canada) to estimate dipole location. In this article, the spatial origin of an IC is approximated with the location of its associated dipole.

We opted for the BeMoBIL pipeline after comparing it against the APP pipeline (da Cruz et al., 2018), which proved to be less robust for our dataset. We based this conclusion on different metrics, by evaluating each artifactual detection step (number of channels removed, proportion of time samples excluded) and by assessing the performance of the subsequent ICA (mutual information reduction and remaining pairwise mutual information, Delorme et al., 2012). In particular, the BeMoBIL pipeline proved to be more stable and conservative than the APP pipeline (rejecting more artifactual channels and noisy temporal segments, both more consistently across participants). We detail the comparison and its results in Methods S3 and Figure S4, respectively.

2.7.2 | Individual IC labeling

We used the ILabel algorithm (version 1.1, Pion-Tonachini et al., 2019) with the “default” option to give an automatic class prediction for each IC. The model supporting this algorithm considers seven classes: (1) Brain, (2) Muscle, (3) Eye, (4) Heart, (5) Line Noise, (6) Channel Noise, and (7) Other. The prediction takes the form of a compositional label: a percentage vector expressing the likelihood of the IC to belong to each of the considered classes. Then, it compares each percentage to a class-specific threshold to form the IC label. We used the threshold vector reported by Pion-Tonachini et al. (2019) for optimizing the testing accuracy. Considering the recentness of this algorithm and the fact it has never been validated on mobile EEG data, we refined the labeling process to increase its conservativeness on Brain ICs. After the initial categorization by the algorithm, we automatically examined the ECD of ICs passing the Brain threshold and we rejected all ICs whose ECD was either located outside brain volume or exhibiting residual variance over 15% (commonly accepted threshold for dipolarity, see Delorme et al., 2012) and we put them in the “Other” class. Residual variance quantifies the quality of the fit between the actual topographic activation map and the estimated dipole projection on the scalp. Among the remaining ones, we distinguished two cases: (1) if the IC label was uniquely “Brain”, we automatically accepted it; (2) if the IC label was hybrid (multiple classes above threshold), we manually inspected the IC properties to assign the label ourselves according to the ILabel guidelines (https://labeling.ucsd.edu/tutorial/labels – an example can be found in Figure S3). To all ICs below brain threshold, we assigned unique labels based on their highest percentage class.
2.8 | Group-level EEG analysis

In order to retain maximal information for further processing, for each participant we copied the ICA results (decomposition weights, dipole locations, and labels) back to the continuous version of the dataset (i.e., the dataset before time domain cleaning in the BeMoBIL pipeline). We first bandpass filtered the data between 1 Hz (zero-phase Hamming windowed finite impulse response high-pass filter with 0.5 Hz cut-off frequency and 1 Hz transition bandwidth) and 40 Hz (zero-phase Hamming windowed finite impulse response low-pass filter with 45 Hz cut-off frequency and 10 Hz transition bandwidth). We then epoched each dataset into trials, starting at the beginning of the baseline period and ending at the time of trial completion. For each IC and each trial, we computed the trial spectrum using the pwelch function of EEGLAB (1 to 40 Hz in linear scale, using a wavelet transformation with three cycles for the lowest frequency and a linear increase with frequency of 0.5 cycles). Using a gain model, we individually normalized each trial with its average over time (Grandchamp & Delorme, 2011). Separately for each participant, we calculated a common baseline from the average of trial baseline periods (condition specific) and we subsequently corrected each trial with the baseline corresponding to its experimental condition (gain model). At the end, power data were log-transformed and expressed in decibels. To enable trial comparability, these event-related spectral perturbations (ERSPs) were time-warped based on the same sequence of events as for the zone-based analysis.

2.8.1 | Component clustering

To allow for a group-level comparison of EEG data at the source level (ICs), we selected the 70 first ICs outputted by the AMICA algorithm, which corresponded to ICs explaining most of the variance in the dataset (Gramann et al., 2018). This ensured the conservation of 90.6 ± 1.8% (mean ± SE) of the total variance in the dataset while greatly reducing computational cost and mainly excluding ICs with uncatagorizable patterns. We conducted this selection independently of the class label for each IC. We applied the repetitive clustering region of interest (ROI) driven approach described in Gramann et al. (2018). We tested multiple sets of parameters to opt for the most robust approach and we present here the selected one (the detailed procedure for this comparison can be found in Methods S4 and its results in Table S3). We represented each IC with a 10-dimensional feature vector based on the scalp topography (weight = 1), mean log spectrum (weight = 1), grand average ERSP (weight = 3), and ECD location (weight = 6). We compressed the IC measures to the 10 most distinctive features using PCA. We repeated the clustering 10,000 times to ensure replicability. According to the results from parameters comparison (see Methods S4), we set the total number of clusters to 50 and the threshold for outlier detection to 3 SD in the k-means algorithm. This number of clusters was chosen inferior to the number of ICs per participant to favor the analysis of clusters potentially regrouping ICs from a larger share of participants and therefore more representative of our population. We defined [0, −55, 15] as the coordinates for our ROI, a position in the anatomical region corresponding to the retrosplenial cortex (BA29/BA30). We set the first coordinate (x) to 0 because we did not have any expectation for lateralization. Coordinates are expressed in MNI format. We scored the clustering solutions following the procedure described in Gramann et al. (2018). For each of the 10,000 clustering solutions, we first identified the cluster whose centroid was closest to the target ROI. Then, we inspected it using six metrics representative of the important properties this cluster should fulfill (see Table S3 detailing the comparison procedure results). In order to combine these metrics into a single score using a weighted sum (same weights used to choose the best of the 10,000 solutions), we linearly scaled each metric value between 0 and 1. We eventually ranked the clustering solutions according to their score and selected the highest rank solution for the subsequent data analysis.

2.8.2 | Cluster labeling

We then inspected the 50 clusters given by the selected clustering solution. We first used the individual IC class labels to compute the proportion of each class in the clusters. As the clustering algorithm was blind to the individual class labels, most clusters contained ICs with heterogeneous labels. Bearing in mind that the ICLabel algorithm has not been validated on mobile EEG data yet, we suspected that the observed heterogeneity could, to a certain extent, owe to individual labeling mistakes. We therefore performed a manual check (identical to the hybrid case in the Individual IC labeling section above) of individual IC labels in specific clusters exhibiting a potential interest for the analysis. These clusters were those with at least 20% of Brain label, those with at least 50% of Eye label, and those located in the neck region with at least 50% of Muscle label. Indeed, both eye and muscle activity are inherent to the nature of the mobile EEG recordings and their...
analysis can inform us on participants' behavior (Gramann et al., 2014), similarly to horizontal head rotations and landmark visibility variables, with a finer temporal resolution. We finally labeled every cluster from their most represented class after correction, only when this proportion was above 50%. Eventually, within each of the labeled clusters, we removed the ICs whose label did not coincide with the cluster label.

2.8.3 | Clusters analysis

We computed single-trial ERSPs as for the clustering procedure. To get the cluster-level ERSP, we took the arithmetic mean of the power data first at the IC level (including the baseline correction), then at the participant level, and finally at the behavioral group level. At the end of these operations, we log-transformed the power data to present results in decibels. We performed statistical analysis comparing ERSP activity between trial type (learning, control, probe), using a non-parametric paired permutation test based on maximum cluster-level statistic (Maris & Oostenveld, 2007) with 1,000 permutations. For each permutation, we computed the \( F \)-value for each “pixel” (representing spectral power at a given time–frequency pair) with an \( 1 \times 3 \) ANOVA. As the ANOVA test is parametric, we used log-transformed data for statistical analysis as ERSP sample distribution has a better accordance with Gaussian distribution in that space (Grandchamp & Delorme, 2011). We selected samples with \( F \)-value above 95th quantile of the cumulative \( F \)-distribution and clustered them by neighborhood. The cluster-level \( F \)-value was the cumulative \( F \)-value of all samples in the cluster. We then formed the distribution of observed maximum clustered \( F \)-values across permutations to compute the Monte Carlo \( p \)-value for the original repartition. As a post hoc test, we repeated the same analysis for each pair of conditions, with \( t \)-values instead of \( F \)-values and two-tailed \( t \)-test instead of ANOVA. We finally plotted ERSP differences only showing samples significant for both the three conditions permutation test and the inspected pairwise permutation test. The significance level was \( p < 0.05 \) for all tests in this case.

3 | RESULTS

3.1 | Behavioral results

3.1.1 | Goal-oriented navigation performance

During control trials, all participants successfully solved the Y-maze task by consistently choosing the goal arm (Figure S1a, left). During the probe trials, 14 participants navigated to the correct goal arm (i.e., allocentric response), whereas 2 participants went to the error arm (i.e., egocentric response; Figure S1a, right).

In terms of time to goal, all participants learned rapidly to locate and navigate to the goal position: after the first learning trial, in which goal finding was merely random, the mean time to goal of the allocentric group plateaued at around 6 s (Figure S1b, left). During control trials, the mean time to goal of allocentric participants remained constant and identical to the plateau reached at the end of the learning condition (Figure S1b, middle). In the probe trials, the mean time to goal of the allocentric group increased slightly by ~1 s as compared to the control condition (Figure S1b, right). Overall, the interindividual variability remained very low, reflecting the simplicity of the navigation task.

3.1.2 | Spatial behavior across conditions and maze zones

We sought to characterize the exploratory behavior as a function of the protocol conditions (Condition factor) as well as of the zones in the Y-maze (Zone factor, see Figure 1c). Hereafter, only the analyses on the allocentric group are presented as only two participants adopted an egocentric behavior (expectedly, Bécu et al., 2020b; see Figure S2 for the individual behavior of egocentric participants).

Horizontal head rotations

First, we assessed the searching behavior by quantifying the horizontal head rotations variability (Figure 3a,b). We did not observe any effect of Condition \((F(2;273) = 2.69, p = 0.069)\), whereas we found a significant effect of the Zone on horizontal head rotations variability \((F(6;273) = 8.99, p < 0.00001)\). Post-hoc analysis indicated that horizontal head rotations variability was higher at the beginning of the trajectory in comparison to the center of the maze (staticS1 versus C, \( t(2) = 3.67, p < 0.01 \); mobileS1 versus C, \( t(2) = 5.5, p < 0.00001 \)). There was no interaction between Condition and Zone for this metric \((F(12;273) = 0.13, p = 0.99)\).

Walking speed

Second, we analyzed the walking speed across different conditions and zones (Figure 3c,d). We found a significant main effect of Zone \((F(6;273) = 472.15, p < 0.00001)\). Post-hoc analysis revealed that the participants spent more time, and exhibited a slower walking speed at the beginning of the starting arm (i.e., in zone S1, both before and after walking onset, \( t(2) < -15, p < 0.00001 \), for all pairwise comparisons involving either staticS1 or mobileS1). We observed a tendential, but
FIGURE 3  Behavioral metrics – Walking speed, horizontal head rotations variability, and landmark visibility for the allocentric group. (a) Average standard deviation of horizontal head rotations, computed from the difference between head and torso orientation. (b) Main effect of Zone on horizontal head rotations variability $F(6;273) = 8.99, p < 0.00001$. (c) Average instantaneous walking speed. (d) Main effect of Zone on walking speed $F(6;273) = 472.15, p < 0.00001$. (e) Average landmark visibility. The color code corresponds to the percentage of time each landmark was visible at the screen. (f) Three-way interaction effect of Zone, Condition, and Landmark on landmark visibility. Each bar shows average landmark visibility (sorted in descending order) for a specific combination of zone (labeled), condition (color), and landmark (texture). We present only combinations associated with at least 10% landmark visibility (17 combinations out of 63). (a, c, e) We divided each trial according to the same sequence of events: walking onset, followed by the first passage in the starting arm (S) then in the finish arm (F), being either the goal or the error arm. Events are horizontally spaced according to the median duration between each event. All three plots represent data in the learning, control, and probe conditions, averaged between separating events across all trials and blocks for all 14 allocentric participants. (b, d, f) Mean value with standard error of the mean (black bars). We present the summary of the significant differences (green braces) found in post-hoc analysis (computed on a pairwise basis, then grouped when similar). For figure (f), we found no pairwise significant differences within the group of combinations not shown (below 10%). ***$p < 0.00001$, **$p < 0.001$, *$p < 0.01$
not significant, main effect of Condition on the walking speed \((F(2;273) = 3.91, p = 0.021)\), which did not survive the multiple comparisons correction). There was no interaction effect between Condition and Zone \((F(12;273) = 0.45, p = 0.94)\).

**Landmark visibility**

Third, we tested the visibility of the landmarks depending on the condition, zone, and landmark (Figure 3e) and we observed a three-way interaction between all factors \((F(24;819) = 25.31, p < 0.00001)\). Post-hoc analysis (Figure 3f) revealed a clear tendency for landmarks being visible in the starting arm of the maze (as opposed to the center zone and the finish arm), modulated by the condition and the landmark attended. Figure 3f shows the landmark visibility of \([\text{Condition}; \text{Zone}; \text{Landmark}]\) combinations in descending order, and we can notice steps of combination triplets with the same Zone factor (from staticS1 to C only), in the order in which they are visited by the participants.

The consistent pattern in each triplet shows a preferred landmark for each condition: Landmark 1 for learning and control trials, and Landmark 2 for probe trials. A slight deviation from the dominant pattern is that the mean visibility of Landmarks 1 & 2 in Center zone during probe trials is found at the same level (Figure 3f), although not statistically different from the visibility of any landmark in any condition in the same zone. All additional statistical results (main effects, two-way interactions) are presented in Table S2.

### 3.2 | EEG cluster analysis

#### 3.2.1 | Independent component selection

To give an overview of the IC inspection and selection process, we provide IC and cluster counts at different steps of our procedure (see Figure 2). In total, we extracted 1,943 ICs of the whole dataset (16 participants). First, at the individual IC labeling step, we relabeled 204 of 394 ICs initially labeled as “Brain” (i.e., automatic rejection based on RV threshold and manual inspection of hybrid cases). Starting from 1,120 input ICs, the clustering algorithm placed 1,047 ICs in valid clusters (73 outliers). Then, to complete the cluster labeling step, we selected 35 “clusters of interest” out of the 50 output clusters. We reviewed 755 ICs and edited the label in 207 of them. Eventually, we removed a total of 414 ICs in disagree-
the anterior cingulate (Cluster 4: [-2,9,22]), BA6 in the right precentral gyrus (Cluster 5: [33,−9,52]), and BA3 in the left postcentral gyrus (Cluster 6: [−37,−28,49]). These coordinates ([x,y,z]) are in Talairach units.

### 3.2.3 | Brain cluster activity

The analyses of the 6 selected brain clusters are presented in Figures 6 and 7 (clusters 1–3 and 4–6, respectively).
Alpha band activity (8–12 Hz)
The average ERSP analysis for posterior parieto-occipital clusters (1–3) showed a marked alpha (8–12 Hz) desynchronization (power suppression of 3 dB or more) starting after trial onset in all conditions (Figure 6b,e,h). This power suppression slowly faded away or narrowed down around 9 Hz when the participant left the first section of the maze. The desynchronization was less marked in the precentral and the postcentral gyri, but it was sustained throughout the trial, except for the control condition (significant difference found for the precentral gyrus near the central zone of the maze, see Figure 7f). In the anterior cingulate, we intermittently observed a similar but reduced alpha power suppression (difference of 1 dB with respect to baseline, Figure 7b).

Gamma band activity (>30 Hz)
We found that gamma (>30 Hz) synchronization was strongly enhanced in this navigation task, with clusters 1–4 (posterior and anterior cingulate, cuneus, and supramarginal gyrus) presenting amplitudes greater than baseline in this frequency band, consistently throughout the maze (Figures 6 and 7). Nonetheless, a power increase in this frequency band was found between trial start and the center, especially in the probe condition (significant differences found in the mobileS1 zone for cuneus, supramarginal gyrus, and anterior cingulate clusters). A comparison between conditions also demonstrated a reduced gamma activity upon reaching the center in the control condition in the posterior cingulate and an increased gamma power in the learning condition in the cuneus in the finish arm.

Delta and theta band activity (<8 Hz)
Finally, we observed modulations of low-frequency rhythms (delta range 1–4 Hz, theta range 4–8 Hz), with sustained greater delta amplitudes in the starting arm in all brain clusters and a strong transient theta burst at the beginning of the trial in posterior parieto-occipital clusters (1–3). The brain activity in these frequency bands proved to be condition specific for these clusters, with a generally higher power for the learning condition along the finish arm (Figure 6c,f,i).
4 | DISCUSSION

This work brings together the technology and data analysis tools to perform simultaneous brain/body imaging during landmark-based navigation in fully mobile participants. Our behavioral results show that a majority of young adults can rapidly learn to solve the Y-maze by using an allocentric strategy, confirming previous findings in similar landmark-based navigation studies (Bécu et al., 2020b; Kimura et al., 2019). We find that allocentric participants have the capacity to flexibly reorient by observing landmarks at the beginning of the trial (consistent
with the more precise gaze dynamics described in Bécu et al., 2020b). The analysis of high-density EEG data shows that exploitable neural signals are extracted from various brain regions (posterior cingulate, cuneus/precuneus, supramarginal gyrus, anterior cingulate, precentral gyrus, and postcentral gyrus) that replicate and extend previous neuroimaging findings from a similar fMRI study (Ramanouë et al., 2020). Overall, the identified brain structures represent an extended ensemble of areas involved in the high-level processing of visual information, in spatial representation, and in motor planning necessary to navigate.

FIGURE 7  Detailed analysis of brain clusters 4–6 for the allocentric group. The layout is the same as in Figure 6a,d,g Topographical map of the average cluster components’ projection at the scalp level (top) and sagittal/frontal views of all ICs in the cluster (bottom). (b, e, h) ERSP average per condition. (c, f, i) ERSP pairwise differences between conditions. “L > C”: difference between Learning and Control, “P > C”: difference between Probe and Control, “P > L”: difference between Probe and Learning. (a-c) Cluster 4 – Anterior Cingulate. In the allocentric group, this cluster contains 14 ICs from 11 different participants. (d-f) Cluster 5 – Right Precentral Gyrus. In the allocentric group, this cluster contains 15 ICs from 11 different participants. (g-i) Cluster 6 – Left Postcentral Gyrus. In the allocentric group, this cluster contains 12 ICs from 10 different participants.
4.1 Task-solving behavior

The zone-based analysis reveals a common task-solving behavior in allocentric participants, starting by an observation period at the beginning of the trial (slow speed, high variability in horizontal heading, and maximal visibility of landmarks) followed by navigation to the chosen arm. We found minimal head rotations after reaching the center of the maze and no significant deceleration, indicating the initial observation period to be the main source of visual information for the participants. This interpretation was also supported by the analysis of horizontal eye movement and neck muscle activity clusters, which exhibited greater activity at the beginning of the maze (Figure 4b,e,h). Interestingly, between-condition contrasts revealed an accentuation of this pattern in the probe condition (Figure 4c,f), probably reflecting a higher need for information gathering at an unfamiliar starting point.

The landmark visibility analysis confirmed that navigators may rely extensively on the landmark appearing straight ahead at the beginning of the task, which differs in the probe condition (Landmark 2) from the other conditions (Landmark 1). This suggests that the participants were mainly capable of reorienting with the information gathered from this landmark only, even in the probe condition. We found additional tendencies for this metric (see last columns of Figure 3f), such as (a) a similar visibility for the two predominant landmarks at the center during probe condition and (b) an increased visibility for the second most visible landmark (relative to each condition) with respect to the third during the initial observation period in the learning and probe conditions. None of these observations were applicable to the two egocentric participants (Figure S2e,f). Therefore, these findings might reflect a perceptual mechanism helping to bind multiple landmarks in a single representation of the environment, specific to the allocentric participants’ strategy.

4.2 Anatomical substrates of the clusters retrieved

We hypothesized that the analysis of brain dynamics in fully mobile individuals would retrieve structures involved in active, multimodal landmark-based spatial navigation. Here, we contrast our results against those from static neuroimaging paradigms. First, we expected the retrosplenial complex (RSC) to play a central role in solving the Y-maze, as this task requires landmark-based reorientation. Accordingly, we found a cluster located in or near the posterior cingulate cortex (cluster 1), encompassed by the RSC (Julian et al., 2018). fMRI studies have consistently shown that the human RSC encodes heading direction (Marchette et al., 2014; Shine et al., 2016) anchored to local visual cues, like stable landmarks, by using a first-person perspective (Auger & Maguire, 2018; Auger et al., 2012; Marchette et al., 2015). Moreover, the RSC is embedded in a network of somatosensory areas that were also partially retrieved in our cluster analysis. In particular, the precuneus (cluster 2) is involved in several aspects of spatial cognition, such spatial attention, spatial working, long-term memory, and the representation of landmarks, in association with RSC during navigation (Cona & Scarpazza, 2019). Also, in line with the mobile aspects of our paradigm, several studies have highlighted the role of the precuneus in the integration and coordination of motor behavior during navigation (Navarro et al., 2018). Around the centroid of the cluster associated with the precuneus, we note that the ICs forming the cluster are distributed across the anatomical boundaries between occipital and parietal cortices. This region spans areas mediating visuo-spatial processing, such as the Occipital Place Area, which is sensitive to navigable pathways in a perceived scene (Bonner & Epstein, 2017; Patai & Spiers, 2017).

Our EEG analysis also retrieved three clusters associated with the supramarginal gyrus (cluster 3), the anterior cingulate cortex (cluster 4), and the precentral gyrus (cluster 5). The supramarginal gyrus, which belongs to the somatosensory cortex, plays a role in the mnemonic components of spatial navigation (Sneider et al., 2018; van der Linden et al., 2017). In addition, as it is encompassed within the inferior parietal lobule, the supramarginal gyrus is involved in spatial attention (Cona & Scarpazza, 2019). The anterior cingulate cortex suberves high-level cognitive functions such as route planning (Lin et al., 2015; Spiers & Maguire, 2006) as well as its re-evaluation and updating based on internal monitoring, more specifically with respect to error detection and spatial reorientation (Javadi et al., 2019). As for the right precentral gyrus cluster, we report its centroid in BA6 although the spatial extent of its ICs spans toward more frontal areas. Considering this limited spatial precision, we speculate a putative contribution from the supplementary motor area proper and the right middle frontal gyrus, both overlapping with BA6. The former is recruited in motor planning (Simon et al., 2002) and motor execution of self-initiated movements (Cona & Semenza, 2017), consistent with the mobile aspect of our task. The latter is involved in spatial attention and spatial working memory, specifically in BA6 (Cona & Scarpazza, 2019). Finally, our brain analysis retrieved post-central gyrus activity (cluster 6), encompassing the primary somatosensory cortex, and thus most likely to be involved in the processing of proprioception (Rausch et al., 1998).

4.3 Functional analysis of the clusters’ activity

4.3.1 Gamma band activity (>30 Hz)

An objective of this work was to couple brain and body imaging during spatial navigation. Complementing the behavioral
findings, the analysis of transient time–frequency EEG patterns shows a strong gamma band synchronization in posterior parieto-occipital clusters, especially in the starting arm (Figure 6b,e,h), which coincides with the increased eye movement related activity observed in the same spatial area (low frequencies in Figure 4b). In line with findings showing that increased gamma power in parieto-occipital region promotes sharper visuo-spatial attention (Gruber et al., 1999; Jensen et al., 2007; Müller et al., 2000), this cortical activity pattern supports our interpretation of the participants’ behavior. We reported significant differences when the participant starts walking in the probe condition compared to the learning and control conditions (in the cuneus/precuneus, Figure 6f, and supramarginal gyrus, Figure 6i). This pattern may reflect a greater attentional demand triggered by the visual conflict between the probe and the other conditions, forcing the participant to actively reorient. Statistical analyses conducted on eye and muscle clusters also revealed a more active state (more frequent eye movements and increased muscle activity) when the participant starts walking during probe trials. The greater involvement of posterior parietal cortex (especially the precuneus) during this crucial reorientation moment is coherent with the fMRI evidence linking it to the navigationally relevant representation of landmarks, when participants are moving with respect to stable objects (Cona & Scarpazza, 2019). The mobile EEG literature of locomotion control more often reports activity bound to steady-state gait cycle events (e.g., Castermans et al., 2014; Gwin et al., 2011; Luu et al., 2017b; Wagner et al., 2014, 2016), making it difficult to compare with our experimental design. Several works presenting results contrasting a walking condition with a standing baseline condition described a desynchronization in the high beta band (25–35 Hz) in the sensorimotor cortex (Seeber et al., 2014, 2015; Wagner et al., 2012), which does not concur with our findings (Figure 7e,h). Nonetheless, Bulea et al. (2015) report high gamma band synchronization (30–50 Hz) when comparing active walking to quiet standing in the posterior parietal area, which better aligns with our results (Figure 6b,e,h). However, locomotor control can only be a part of the interpretation as it does not explain the specific activity observed in the probe condition.

4.3.2 | Alpha band activity (8–12 Hz)

Our ERSP analysis shows a desynchronization in the alpha band, spanning almost all clusters (except the anterior cingulate), with different temporal dynamics. In the sensorimotor cortex (post-central and pre-central gyri), the desynchronization extends to the low beta band, it starts a few moments before movement onset and it is sustained throughout the whole maze traversal (Figure 7e,h). This pattern advocates for a mere signature of locomotion, as reported in numerous mobile EEG studies comparing walking and standing (Bulea et al., 2015; Presacco et al., 2011; Seeber et al., 2014). Further supporting the idea that such activity is a neural correlate of ambulation, no differences between conditions were found in postcentral gyrus (Figure 7i). Interestingly, the precentral gyrus exhibited a modulation of activity around the center of the maze where alpha desynchronization was less pronounced in the control condition as compared to others (Figure 7f). The precentral gyrus is known to be associated with movement planning (Navarro et al., 2018; Wagner et al., 2014). Alpha power suppression has been linked to increased activity in motor regions (Pfurtscheller & Klimesch, 1991), such that this activity pattern could reflect a more passive execution of the turn in a situation in which the participant can straightforwardly repeat the learning condition. Although this purely ambulatory feature extends to more posterior parietal areas (Bulea et al., 2015), the temporal dynamics of the alpha power in our parieto-occipital clusters (i.e., an almost immediate desynchronization after trial start and a subsequent fading across maze traversal, Figure 6b,e,h) suggests a different interpretation. According to the meta-analysis from Cona and Scarpazza (2019), the precuneus and the inferior parietal lobule are embedded in a fronto-parietal network mediating spatial attention. Thus, the fading of the desynchronization might reflect a progressive decrease in spatial attention, as sufficient visual information is being gathered. As participants seem to make their decision early in the task, they should reach their maximal degree of alertness in the first sections of the maze and let it drop afterwards. Echoing this interpretation, several EEG studies of spatial navigation associated alpha power in the parietal cortex to spatial learning (Gramann et al., 2010; Lin et al., 2015), with significant task-related modulations.

In an experiment reporting the modulation of RSC activity in passive simulated navigation, those participants who relied on an allocentric reference frame demonstrated a sustained alpha power decrease during straight segments and a strong alpha power increase during absolute heading rotation (Chiu et al., 2012; Lin et al., 2015, 2018). For our posterior cingulate cluster, such heading discriminant activity is neither observed at the starting position where head movements are maximal (relative to the body) nor near the central zone of the maze (relative to the global environment, i.e., the landmarks) (Figure 6b). Partially explaining these diverging results, Gomez et al. (2014) reported a stronger RSC activation during on-the-spot rotation as compared to continuous movement, tempering the heading computation role of RSC when translational movements are involved. Additionally, the alpha desynchronization elicited by a desktop-based rotation is absent when performed physically (Gramann et al., 2018), which shows the important influence of vestibular and proprioceptive cues in modulating RSC activity. Thus, assuming that our posterior cingulate cluster is bound to RSC activity,
our results provide additional evidence that the involvement of RSC in heading calculation has been overestimated with respect to ecological navigation. The dynamics of the posterior cingulate cluster in the alpha frequency band are more in accordance with a memory role serving the encoding/retrieval of the egocentric percepts into the allocentric representation (Mitchell et al., 2018; Vann et al., 2009). Indeed, the fact that alpha desynchronization occurred during the observation of the environment suggests the association of RSC with the encoding/decoding of landmark-based information. This interpretation fits with the tendency of our cluster to be localized in the dorsal part of the posterior cingulate, which is known to play a role in spatial recall tasks in opposition to the ventral/medial part, more likely to be activated during tasks proposing passive viewing or active navigation without the need to respond, perform spatial computation, or self-localize (Burles et al., 2018).

4.3.3 | Delta and theta band activity (<8 Hz)

We observe a strong delta band synchronization (<5 Hz) in all clusters, lasting the whole traversal of the maze, which has been previously reported as a motion-related artifact (Castermans et al., 2014; Gwin et al., 2010; Presacco et al., 2011). However, the presence of condition-specific modulations in the posterior parieto-occipital clusters casts a doubt on this interpretation. During the learning condition, statistical analyses demonstrated a sustained delta/theta synchronization in the finish arm (starting in the center zone, Figure 6c,f,i). Possibly elucidating this feature, a previous study of spatial working memory in mobile conditions observed a similar theta synchronization seconds prior to the stimulus presentation in posterior cingulate and somatosensory association areas (Kline et al., 2014). Arguing that theta power modulations can be related to memory encoding and maintenance, this may be the signature of a learning mechanism, preparing to encode the outcome of the learning trial at the end of the finish arm. However, unlike Kline et al. (2014), we do not find subsequent theta desynchronization on stimulus presentation (goal appearance). Another deviation from the artificial hypothesis is that the delta/theta synchronization seems specific to the starting arm for the anterior cingulate cortex (Figure 7b), coinciding with participants’ decision-making period as indicated by behavioral analyses. This may reflect the increased spatial working memory demand required for route planning, as previous studies reported increased theta power in the frontal cortex during more cognitively demanding navigation periods (Caplan et al., 2003; Kahana et al., 1999; Lin et al., 2015). Closer to the interrogations posed by the present task, Ferguson et al. (2019) found the anterior cingulate to mediate a reinforcement learning role by eliciting a reward when allocentric navigators were shown previously learned cues predicting the goal location. In addition, we observed theta bursts of activity (4–8 Hz) closely time-locked to the beginning of the task in most clusters (mainly in the posterior cingulate, Figure 6b, and supramarginal gyrus, Figure 6h). This pattern of activity may be framed within a postural control interpretation: as the environment suddenly appears to the participant, his/her balance control system, previously deprived of any visual information, needs to be updated based on the novel visual cues (Flückiger & Baumberger, 1988; Horak & Macpherson, 2011). Strikingly, theta bursts of activity were similarly described immediately following spontaneous loss of balance from walking on a beam (Sipp et al., 2013) and sudden visual perturbations to standing or walking balance (Peterson & Ferris, 2018). These bursts were noticeable in posterior cingulate and posterior parietal areas, associated with vestibular sensing (Kim et al., 2017) and resolving visual conflicts (Peterson & Ferris, 2018), respectively.

4.4 | Limitations

Source reconstruction was performed using an electrodes’ location template and average MRI anatomical data, which limited its spatial accuracy. Thus, the interpretations proposed in this work should be treated with caution. The use of subject-specific data to build the head model would help increase the accuracy of source localization algorithms (Akalin Acar & Makeig, 2013; Shirazi & Huang, 2019) and it would eventually enable more robust interpretations of the neural correlates of spatial behavior.

Although the methods employed here to clean the EEG signals have been previously validated in the literature (Nordin et al., 2019; Richer et al., 2019), there is never complete guarantee that the results are artifact free. In particular, the muscular activity associated with microsaccades has been shown to resist standard cleaning methods (Hassler et al., 2011; Yuval-Greenberg et al., 2008) and it could in principle be contributing to the brain ICs in the gamma frequency band (Yuval-Greenberg et al., 2008). However, the influence of this type of artifact is meaningful in experimental setups favoring the accumulation of microsaccades at a fixed latency with respect to the synchronizing event (e.g., fixation of a visual target; Yuval-Greenberg et al., 2008) and the probability that this applies to our setup is low.

Gait-related artifact contamination is another well-known pitfall of ambulatory studies (Castermans et al., 2014). Walking induces small motions of electrodes and cables that can have a large impact on the signal-to-noise ratio. Typically, the spectral signature of such artifacts contains elevated power amplitudes at the stepping frequency (between 0.5 and 1 Hz for normal walking speeds) and its harmonics, as well as a power modulation pattern time-locked to gait cycle
especially marked below 20 Hz (Castermans et al., 2014; Snyder et al., 2015). Therefore, as already acknowledged, the low frequencies (between 1 and 5 Hz) power increase observed consistently through our clusters may reflect this type of contamination. Yet, motion artifacts were found negligible during treadmill walking at moderate speeds such as those adopted by participants in our study (Gwin et al., 2010; Nathan & Contreras-Vidal, 2016). Also, the wireless property of our EEG system (as in Nathan & Contreras-Vidal, 2016) provides additional robustness to gait-related artifacts by minimizing cable sways, identified as major artificial causes (Symeonidou et al., 2018).

Considering the recent publication of the ICLabel algorithm, our work provides some practical insights on how to integrate this promising tool into the MoBI approach. At first, the automatic IC classification has proven to be very useful to deal with large numbers of components. However, through the manual inspection of a large proportion of the automatically assigned labels, we uncovered and corrected a substantial amount of discrepancies between the algorithm and the experimenter’s opinion, thus adopting a semi-automated procedure. Even if human categorization of ICs can be variable and error prone (Pion-Tonachini et al., 2019), we believe that these discrepancies also stem from complex artifact patterns present in mobile EEG. However, resorting to the experimenter’s judgment is not desirable for future studies as it impairs replicability and it is very time-consuming for high-perimenter’s judgment is not desirable for future studies as it impairs replicability and it is very time-consuming for high-density recordings. Future works should explore the flexibility of interpretation offered by the compositional label, for example, by adapting the probability thresholds to better tailor the algorithm’s output to the characteristics of the data being processed.

5 | CONCLUSION & FUTURE WORKS

This study provides a proof-of-concept about the possibility of imaging the neural bases of landmark-based spatial navigation in mobile, ecological set-ups. First, the presented EEG analysis identifies a set of brain structures also found in fMRI studies of landmark-based spatial cognition. Second, our approach reveals the role of brain areas involved in active, fully engaging spatial behavior (such as clusters in the sensorimotor cortex related to motor execution and proprioception), whose contribution is usually overlooked in static fMRI paradigms. We present new insights onto the cortical activity mediating successful spatial reorientation when visual, proprioceptive, and vestibular sensory inputs are coherent. Specifically, alpha band desynchronization in the posterior cingulate when participants gather visual information provides further support to the idea that RSC plays an important role at the interface between perception of landmarks and spatial representation. Despite showing few effects of experimental condition, our results illustrate the benefit, in terms of deciphering neural dynamics within the course of a trial, of fine temporal resolution brain imaging paired with meaningful behavioral markers during spatial navigation.

The methodology associated with the MoBI approach remains quite new and such experiments help to identify vectors of improvement. At the preprocessing stage, further characterization of the parameters and robustness comparison with other pipelines (such as Automatic, Pedroni et al., 2019) would be beneficial. Complementary steps such as sliding window approaches for isolating transient artifacts using principal component analysis and/or canonical correlation analysis can improve source separation compared to ICA alone (Artoni et al., 2017; Nordin et al., 2020). Adding simultaneous noise and neck electromyographic recordings have also been shown to successfully assist the identification and removal of motion-related artifacts (Nordin et al., 2019, 2020). Concerning the gathering of insights on strategy-specific behaviors, additional improvements of the protocol are also desirable. Using a passively guided traversal of the maze as a baseline to contrast with the main task may help to disentangle the neural correlates of locomotion control and active landmark-based spatial navigation. The addition of an eye-tracking system embedded in the VR head-mounted display would also bring further insights on the differential role of visuo-spatial cues (Bécu et al., 2020a).

ACKNOWLEDGMENTS

We first thank the participants of this study for their patience and precious contribution. We thank Elisa Tartaglia for the helpful discussions along the project, and Benjamin Paulisch for his great help in setting up the experiment. We also deeply thank Marion Durteste for her text proof-reading. This research was supported by the Chair SILVERSIGHT ANR-18-CHIN-0002, the LabEx LIFESENSES (ANR-10-LABX-65), and the IHU FOReSIGHT ANR-18-IAHU-01.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

AD, JBDSA, SR, MB, RC, KG, and AA designed the experiment. AD and JBDSA conducted the experiment and collected the data. AD, JBDSA, LG, and MK analyzed the data. AD, JBDSA, SR, and AA wrote the manuscript. AD, JBDSA, SR, MB, LG, MK, RC, JAS, KG, and AA revised the manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ejn.15190.
DATA AVAILABILITY STATEMENT
Data and scripts used for the analysis are available by request to the corresponding authors (Alexandre Delaux, alexandre.delaux@inserm.fr; Jean-Baptiste de Saint Aubert, jean-baptiste.de-saint-aubert@inserm.fr).

ORCID
Alexandre Delaux https://orcid.org/0000-0002-4553-9371
Jean-Baptiste de Saint Aubert https://orcid.org/0000-0001-8304-4989
Stephen Ramanooli https://orcid.org/0000-0003-4733-1097
Marcia Bécu https://orcid.org/0000-0003-4564-1023
Lukas Gehrke https://orcid.org/0000-0003-3661-1973
Marius Klug https://orcid.org/0000-0001-8667-3457
Ricardo Chavarriaga https://orcid.org/0000-0002-8879-2860
José-Alain Sahel https://orcid.org/0000-0002-4831-1153
Klaus Gramann https://orcid.org/0000-0003-2673-1832
Angelo Arleo https://orcid.org/0000-0002-1990-1971

REFERENCES
Cona, G., & Semenza, C. (2017). Supplementary motor area as key structure for domain-general sequence processing: A unified


Iaria, G., Petrides, M., Dagher, A., Pike, B., & Bohbot, V. D. (2003). Cognitive strategies dependent on the hippocampus and caudate...


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.

---

Invariability of EEG error-related potentials during continuous feedback protocols elicited by erroneous actions at predicted or unpredicted states

Fumiaki Iwane1,2,∗, Iñaki Iturrate3,4, Ricardo Chavarriaga3,5 and José del R Millán2,3,6,∗

1 Learning Algorithms and Systems Laboratory (LASA), École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
2 Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, United States of America
3 École Polytechnique Fédérale de Lausanne (EPFL), Campus Biotech, 1202 Genève, Switzerland
4 Amazon, Barcelona, Spain
5 ZHAW Datalab, Zurich University of Applied Sciences, Winterthur, Switzerland
6 Department of Neurology, The University of Texas at Austin, Austin, TX 78712, United States of America

∗ Authors to whom any correspondence should be addressed.
E-mail: fumiaki.iwane@epfl.ch and jose.millan@austin.utexas.edu

Keywords: error-related potentials, continuous feedback protocols, asynchronous decoding, transfer learning

Abstract

Objective. When humans perceive an erroneous action, an EEG error-related potential (ErrP) is elicited as a neural response. ErrPs have been largely investigated in discrete feedback protocols, where actions are executed at discrete steps, to enable seamless brain-computer interaction. However, there are only a few studies that investigate ErrPs in continuous feedback protocols. The objective of the present study is to better understand the differences between two types of ErrPs elicited during continuous feedback protocols, where errors may occur either at predicted or unpredicted states. We hypothesize that ErrPs of the unpredicted state is associated with longer latency as it requires higher cognitive workload to evaluate actions compared to the predicted states.

Approach. Participants monitored the trajectory of an autonomous cursor that occasionally made erroneous actions on its way to the target in two conditions, namely, predicted or unpredicted states. After characterizing the ErrP waveform elicited by erroneous actions in the two conditions, we performed single-trial decoding of ErrPs in both synchronous (i.e. time-locked to the onset of the erroneous action) and asynchronous manner. Furthermore, we explored the possibility to transfer decoders built with data of one of the conditions to the other condition.

Main results. As hypothesized, erroneous actions at unpredicted states gave rise to ErrPs with higher latency than erroneous actions at predicted states, a correlate of higher cognitive effort in the former condition. Moreover, ErrP decoders trained in a given condition successfully transferred to the other condition with a slight loss of classification performance. This was the case for synchronous as well as asynchronous ErrP decoding, showing the invariability of ErrPs across conditions.

Significance. These results advance the characterization of ErrPs during continuous feedback protocols, enlarging the potential use of ErrPs during natural operation of brain-controlled devices as it is not necessary to have different decoders for each kind of erroneous conditions.

1. Introduction

Brain-computer interfaces (BCIs) allow people to establish a new interaction link with the external world by controlling neuroprosthetic devices using mental commands without any physical movement [1–3]. Non-invasive BCIs often measure humans’ spatiotemporal brain activity by using electroencephalogram (EEG), and most approaches rely on the decoding of sensorimotor rhythms [4–6] or so-called P300 event-related potentials [7–9].

© 2021 IOP Publishing Ltd
Lately, BCIs are also incorporating brain responses of human awareness to erroneous actions [10, 11]. The presence of such neuronal correlates has been confirmed by a variety of neuroimaging techniques, including invasive electrocorticography [10, 12], non-invasive functional magnetic resonance imaging [13] and EEG [14]. In EEG, such neuronal correlates are often called error-related potential (ErrP). ErrP is normally characterized by sequential negative (ERN) and positive (Pe) peaks, time-locked to the onset of erroneous actions, and a power increase in the theta band, from 4 to 10 Hz, around the frontocentral area [15–18]. In previous works, detection of ErrPs has been used to correct miss-classified EEG decoding outputs [19–24] or to teach correct policies to external devices [25–29]. These previously investigated ErrP-BCIs were restricted to discrete feedback protocols, where actions are executed at discrete steps. Since many of the target brain-controlled devices (e.g. exoskeletons, wheelchairs and other kind of robots) operate continuously, it is crucial to extend the analysis of ErrPs to such continuous feedback protocols.

Detecting ErrPs in continuous feedback protocols requires to deploy asynchronous decoding techniques, as human operators must monitor continuously for eventual erroneous actions throughout the task. Some previous works have explored the presence of ErrPs in the continuous feedback protocols in interactive experimental environments, where erroneous events corresponded to sudden discrepancies in the execution of commands; we will refer to these cases as errors induced in \textit{unpredicted states} [30–33]. Alternatively, other protocols introduced errors that consistently happened at the same moment while people monitored actions; hereby referred to as \textit{predicted states} [23, 34, 35].

The objective of the present study is to better understand the electrophysiological differences between ErrPs induced in predicted and unpredicted states during continuous feedback protocols. A major reason for such a study is that both types of erroneous conditions, i.e. predicted or unpredicted state, may naturally occur in the same task during continuous feedback protocols. For example, a brain-controlled device could fail either to execute the correct action at a given subgoal in the way to the goal (predicted state) or to reach the precise location of the goal (unpredicted state).

Iturrate et al [34, 36] and Abu-Alqumsan et al [37] showed the variability of the ErrPs across different tasks and different feedback modalities, respectively. Both studies concluded that variation in the ErrP patterns may be attributed to the different cognitive effort required to evaluate each task. Additionally, Yazmir et al [38–40] compared the two types of ErrPs in an experimental protocol with a 3D virtual display where participants or a computer application miss to perform ball hitting actions. They reported that levels of difficulty to predict future actions is a factor that modulates the shapes of ErrPs. Altogether, shapes of ErrPs are modulated by the associated cognitive effort to evaluate one’s action, thus, it is recommended to have task-specific decoders to achieve the highest possible classification performance. We therefore hypothesize that this will be also the case for ErrPs elicited by continuous protocols feedback at predicted or unpredicted states. Specifically, our hypothesis is that ErrPs of unpredicted states would have longer latency than predicted states because of the higher cognitive effort that human operators require to deploy in order to evaluate actions, thus necessitating two different decoders to properly detect ErrPs of each condition. Indeed, in the unpredicted states, they cannot precisely anticipate when erroneous actions may happen, while they can for the predicted states.

Recently, transfer learning approaches have been intensively studied in the field of ErrP-based brain computer interaction in an attempt to reduce calibration time for collecting training data. Several studies have performed transfer learning across different types of errors, from observation ErrPs to interactive ErrPs [41, 42], different applications, from computer application to the humanoid robot interaction [43], and different subjects [44, 45]. They have consistently reported a degradation of classification performance when transferring the decoder from one condition to another.

In the present study, we designed an experimental continuous feedback protocol where an autonomous follower followed predefined trajectories to reach targets. Occasionally the cursor made erroneous movements that prevented achieving the task. In this experiment, ErrPs were elicited in the two different conditions of interest, namely, at predicted states corresponding to subgoals or variable unpredicted states such as the exact location of the final goal. As hypothesized, erroneous actions at unpredicted states gave rise to ErrPs with higher latency than erroneous actions at predicted states, a correlate of higher cognitive effort in the former condition [26, 34, 36, 37]. Nevertheless, and contrary to our expectations, ErrP decoders trained in a combined condition was successfully applied to the each of the two conditions without loss of classification performance. This was the case for both, synchronous as well as asynchronous ErrP decoding. This novel result enlarges the potential use of ErrPs during natural operation of brain-controlled devices as it is not necessary to have task-specific decoders for each kind of erroneous condition.

2. Materials and methods

2.1. Experimental protocol

Eleven healthy participants (nine males and two females, 20 to 30 years old) with normal or corrected-to-normal vision and no prior record of neurological
disorder participated in the experiment. The number of total participants was chosen based on prior studies on ErrPs [26, 29, 35]. We confirmed that all participants had no problem with perceiving the different colors used in the visual interface. The experimental protocol was approved by the local ethics commission (PB_2017-00295). Written informed consent was collected from all participants before starting the experiment. During the experiment, they were asked to sit on a comfortable chair in front of a computer with a 15-inch display.

As shown in figure 1(a), the task was displayed with three concentric circles of different colors, i.e. red, yellow and green, and dashed lines for each \( \pi/4 \) rad. Colors of the concentric circles were chosen to keep the consistency with the traffic light (green, yellow and red). A blue circle represented the cursor, which moved autonomously along the dashed lines. Participants had no control of the trajectory of the cursor. Seven white circles and one black circle in the red area represented erroneous goals and the correct goal, respectively. We designed the circular experimental setup to have a consistent distance between the cross of the center and each concentric circle.

The cursor movement started from the end position of a previous trajectory. Two seconds after the end of a previous trajectory, a new correct goal was randomly selected among seven potential goals (excluding the goal location the closest to the end of the previous trajectory) and highlighted. Once the participant pressed the space key, the cursor started a new trajectory, moving from its current location towards the center along a dashed line. At the center, the cursor made a turn and continued moving along one of the other seven dashed lines. Finally, the trajectory ended when cursor stopped automatically at either the green circle, the yellow circle, at the target on the red circle or outside of the red circle (i.e. overshoot the target). The velocity of the cursor was constant during the experiment. The duration of the cursor movement was 1.5 s between the center and the green circle, 1.8 s between the center and the yellow circle, 2.0 s between the center and the red circle and 2.3 s between the center and outside of the red circle. The average duration of the cursor trajectory was 4.0 s (minimum: 3.0 s, maximum: 4.6 s). The two types of actions, turning at the center and stopping at the final location, were separated 1.5 s apart at least, thus, there was no overlap of the time windows between them.

All participants were instructed to monitor the behavior of the cursor and evaluate both if the cursor was moving towards the correct goal, i.e. angular condition, and stopping on top of the correct goal, i.e. distance condition. Participants were instructed to fix their gaze at the center of the circles to reduce EEG contamination due to ocular artifacts during and until two seconds after the cursor movement. Throughout the experiment, they were asked to remain calm and relaxed and avoid excessive eye blink. The angular condition corresponds to a predicted state in which errors can only occur at a fixed position (i.e. the moment when the cursor passes over the center), while the distance condition represents an
unpredicted state where errors can occur at any position (i.e. stopping before the goal).

Ten runs of 40 cursor trajectories each were recorded for each participant. Three-min breaks were interleaved between each run. For each trajectory, there was 1 angular trial, and eventually 1 distance trial in case of a correct angular trial. In 30% of the trajectories, the cursor followed a direction towards a potential goal different from the correct goal. Within the remaining 70% of the trajectories with correct direction, 70% of them stopped at the correct location of the goal, i.e. the red circle. In order to ensure that participants kept attending during the distance trials, we only used trajectories with the correct angular trials for distance trials. Thus, there was no trajectory which contained both angular and distance errors. On average, there were 277 ± 8 (mean ± std) and 123 ± 8 correct and erroneous angular trials, respectively. Correspondingly, there were 186 ± 7 and 91 ± 7 correct and erroneous distance trials, respectively. The onset of events (or trials) were different for each condition. The onset of an angular trial occurred when the cursor made a turn at the center towards one of the goals. The onset of a distance trial corresponded to the time when the cursor stopped moving.

2.2. EEG processing
EEG signals were recorded throughout the experiment from sixteen active electrodes located at Fz, FC3, FC1, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2 and CP4 (10/10 international system) using a g.USBamp (g.tec medical technologies). The ground electrode was placed on the forehead (AFz) and the reference electrode on the left earlobe. EEG signals were recorded at 512 Hz and power-line notch filtered at 50 Hz by the device. Supplementary figure 1 (available online at stacks.iop.org/JNE/18/046044/mmedia) illustrates the pipeline of the signal processing pipeline for both the time-frequency and classification analysis.

2.2.1. Time-frequency analysis
We firstly investigated event-related spectral perturbation (ERSP) of the induced ErrPs via time-frequency decomposition. To this end, acquired EEG signals were firstly band pass filtered at (1, 40) Hz with a causal 4th order Butterworth filter, and ERSPs were computed within a time window of (−0.5, 1.0) s with respect to the onset of trials. Specifically, the time-frequency power of each single-trial EEG was calculated by the Morlet Wavelet transform [46]. The frequency range was set between 4 and 40 Hz, resulting in a wavelet coefficient matrix with 180 time points and 37 log-spaced frequency bins. Each of the 180 time windows are composed of 427 samples, i.e. 0.83 s; while the number of cycles ranges from 3 to 7 at highest. The resulting coefficients were used to extract spectral power.

In order to extract ERSP, we used the baseline correction method proposed in [47]. This approach is less sensitive to noisy trials than classical baseline correction methods, and produces a non-skewed power distribution. In detail, separately for each participant, condition of actions (angular or distance) and class of actions (correct or erroneous), we applied a single trial full-epoch baseline correction, before averaging across trials and removing the trial-averaged pre-stimulus (i.e. from −0.25 to 0.0 s) baseline. It should be noted that the baselines were corrected using the gain model assumption (i.e. divide by the baseline) as opposed to the additive model (i.e. subtract the baseline) [47], and that the trial-averaged pre-stimulus corrected time frequency power were log-transformed (10log10) for the subsequent statistical analysis.

To extract regions in which time-frequency power was significantly different from one condition to another, we used an non-parametric pixel-based permutation test [48]. Specifically, we performed pixel-based multiple comparison collection (paired t-test, corrected by a 2000 random permutation test, p < 0.05). Comparison of the ERSP was performed for three different pairs, i.e. correct angular against erroneous angular trial, correct distance against erroneous distance trial and erroneous angular against erroneous distance trial.

2.2.2. Synchronous classification
In the synchronous classification, the EEG signals, time-locked to the onset of events, were used to test the hypothesis that neural correlates induced by perception of correct or erroneous actions of the two conditions can be recognized at the single-trial level and that differ from each other. Acquired EEG signals were band pass filtered at (1, 10) Hz with the aforementioned band pass filter and segmented into epochs within a time window of (−1.3, 1.5) s with respect to the onset of trials. The choice of this cut-off frequency was based on previous studies showing the main oscillatory signature of ErrP appears in the theta band [11, 17] and the results of the time-frequency analysis. To further enhance the signal-to-noise ratio (SNR) while reducing the number of source signals, we applied a spatial filter based on canonical correlation analysis (CCA) to EEG signals within the range of (0.2, 1.0) s with respect to the onset [49, 50]. The choice of the time window was determined based on the grand-averaged signals exhibiting electrophysiological difference between the correct and erroneous trials, as shown in figure 3, which is similar to the previous studies performing synchronous classification analysis [26, 35, 51]. The CCA spatial filter method transforms the averaged ErrPs to a subspace containing different ERP components, leading to the increase of SNR [52]. Only the first four components were kept for further analysis.
From these components, we extracted two different types of features, i.e. downsampling the signals and power spectral densities in the theta band [53, 54]. Specifically, temporal signals within the time window of (0.2, 1.0) s with respect to the onset were downsampled to 64 Hz, i.e. 52 features per component, 208 features in total. We used the Welch’s method to estimate the power spectral density (PSD) of each CCA component from 4 to 10 Hz, with a resolution of 2 Hz, in the same time window of (0.2, 1.0) s, 4 features per component, 16 features in total. Based on the training data only, the extracted 224 features (208 temporal and 16 PSD) were then normalized within the scale of (0.0, 1.0) by performing min-max normalization. The normalized features were used to train a classifier based on diagonal linear discriminant analysis (diagonal LDA) that yielded the estimated posterior probability of (0.2, 1.0) s, 4 features per component, 16 features in total. In order to assess the classification performance, we implemented the same signal processing described in section 2.2.2 and obtained a decoder that estimates the posterior probability from an input of 0.8 s of EEG signal. The decoder was applied using a sliding window approach to segmented EEG signal ranging from −1.3 to 1.5 s with respect to the onset of events, what allows to observe the temporal modulation of the posterior probability. Specifically, we used a 0.8 s sliding window with a shift of 1 sample, i.e. 1/512 s to estimate the posterior probabilities in the time period of (−0.5, 1.5) s with respect to the event onset, which corresponds to the EEG signals of (−1.3, 1.5) s. In order to asynchronously monitor the presence or the absence of ErrPs, we determined the optimal classification threshold (nested cross validation, see below). This optimal classification threshold served as a decision boundary to determine if the corresponding posterior probability is correct or erroneous. To this end, the asynchronous classification performance was assessed using the 2 × 2 confusion matrix for each classification threshold \( \tau \in [0, 1] \) with a step of 0.01. Specifically, a correct trial was successfully inferred as correct if the continuous posterior probability did not exceed the optimal classification threshold at any time (true negative trial). Alternatively, an erroneous trial was successfully inferred as erroneous if the continuous posterior probability exceeded the optimal classification threshold within the time window of (0.5, 1.5) s with respect to the onset of the trial (true positive trial). An erroneous trial was successfully inferred as erroneous if the continuous posterior probability exceeded the optimal classification threshold within the time window of (0.5, 1.5) s with respect to the onset of errors. Note that this kind of wrong detection is accounted as a false negative trial. Note also that these are rather stringent criteria, as true negative requires that none of the 1024 outputs of the BCI during the evaluation period exceeds the classification threshold and true positive demands this classification threshold to be surpassed in a limited time window. An optimal classification threshold was determined to be the one which maximized the balanced accuracy (BAC), which is given by \( \frac{1}{2}(TPR + TNR) \). We used this measure as it yielded better classification performance compared to TPR, which was used in the previous study [51]. After computing both the decoder and the optimal classification threshold, these parameters were applied to unseen data to properly evaluate the asynchronous classification performance. The classification performance was measured in terms of BAC.

### 2.2.3. Asynchronous classification

EEG activity was further analyzed asynchronously to estimate when users perceived erroneous actions of the cursor. In order to assess asynchronous classification performance, we implemented the same signal processing described in section 2.2.2 and obtained a decoder that estimates the posterior probability from an input of 0.8 s of EEG signal. The decoder was applied using a sliding window approach to segmented EEG signal ranging from −1.3 to 1.5 s with respect to the onset of events, what allows to observe the temporal modulation of the posterior probability. Specifically, we used a 0.8 s sliding window with a shift of 1 sample, i.e. 1/512 s to estimate the posterior probabilities in the time period of (−0.5, 1.5) s with respect to the event onset, which corresponds to the EEG signals of (−1.3, 1.5) s. In order to asynchronously monitor the presence or the absence of ErrPs, we determined the optimal classification threshold (nested cross validation, see below). This optimal classification threshold served as a decision boundary to determine if the corresponding posterior probability is correct or erroneous. To this end, the asynchronous classification performance was assessed using the 2 × 2 confusion matrix for each classification threshold \( \tau \in [0, 1] \) with a step of 0.01. Specifically, a correct trial was successfully inferred as correct if the continuous posterior probability did not exceed the optimal classification threshold at any time (true negative trial). Alternatively, an erroneous trial was successfully inferred as erroneous if the continuous posterior probability exceeded the optimal classification threshold within the time window of (0.5, 1.5) s with respect to the onset of the trial (true positive trial). An erroneous trial was successfully inferred as erroneous if the continuous posterior probability exceeded the optimal classification threshold within the time window of (0.5, 1.5) s with respect to the onset of errors. Note that this kind of wrong detection is accounted as a false negative trial. Note also that these are rather stringent criteria, as true negative requires that none of the 1024 outputs of the BCI during the evaluation period exceeds the classification threshold and true positive demands this classification threshold to be surpassed in a limited time window. An optimal classification threshold was determined to be the one which maximized the balanced accuracy (BAC), which is given by \( \frac{1}{2}(TPR + TNR) \). We used this measure as it yielded better classification performance compared to TPR, which was used in the previous study [51]. After computing both the decoder and the optimal classification threshold, these parameters were applied to unseen data to properly evaluate the asynchronous classification performance. The classification performance was measured in terms of BAC.
Figure 2. Event-related spectral perturbations (ERSPs) of erroneous trials. (a) Erroneous angular ERSPs within the time window of \((-0.5, 1.0)\) s with respect to onsets of actions within the frequency window of \((4, 40)\) Hz. The region inside the black frame corresponds to the area in which time-frequency power of erroneous angular trials was significantly different from correct angular trials \((\alpha = 0.05)\). (b) Erroneous distance ERSPs within the time window of \((-0.5, 1.0)\) s. No significant difference was observed when compared to correct distance trials nor erroneous angular trials.

Similar to the synchronous classification, we performed asynchronous classification in six different combinations of conditions, corresponding to the same two-by-three matrix as for the synchronous classification. If the training and the test conditions were consistent, ten-outer and five-inner-fold cross validation was performed. Specifically, the inner-fold was used to validate the classification performance for each classification threshold, while the outer-fold was used to train the decoder and to test it together with the validated optimal classification threshold. If the training and the test conditions were inconsistent, the training data was split into five folds to find the optimal threshold, and the decoder was obtained by using the whole training dataset. We kept the same number of folds for the computation of the optimal threshold in order to maintain the consistency across conditions.

Similar to the synchronous classification, we computed the significance threshold \((\alpha = 0.05)\) of the classification performance by performing asynchronous classification with randomly shuffled training labels for 1000 times and the 95% quantiled confidence interval was extracted for each participant.

3. Results

In the following sections, we indicate the pair of training and test datasets as ‘train-test’. For example, Angle-Distance represents the case in which the decoder was trained based on the angular trials, which was then applied to the distance trials.

3.1. Time-frequency analysis

Figure 2 represents the ERSP of erroneous angular trials and erroneous distance trials. In line with previous studies [17, 35], we observed increase of theta band activity 0.2 s after the onset of events \((r = 0)\) for both angular and distance erroneous trials. Such increase of theta band power lasted until 0.8 s with respect to the onsets for both ERSPs. As shown in figure 2(a), we observed a time-frequency region \(((0.25, 0.35) s, (7, 10) Hz)\) in which erroneous angular ERSP was significantly higher than correct angular ERSP \((p < 0.05)\). We did not observe a significant difference for two other comparisons, i.e. correct distance against erroneous distance trials and erroneous angular against erroneous distance trials.

3.2. Electrophysiological results

Figure 3 (a) shows grand averaged signals at the FCz electrode of correct and erroneous trials of the angular condition with respect to the onset of events \((r = 0)\). FCz was chosen for visualization to keep the consistency with previous studies [36, 51], as this electrode has been identified to provide maximal modulation induced by perception of errors [55]. The event related potentials (ERPs) of correct trials exhibited a positive peak at 0.29 s and a negative peak at 0.49 s. The ERP of erroneous trials, i.e. ErrPs, also exhibited the sequential positive, negative peaks at 0.33 and 0.58 s, respectively. Topographical representation of erroneous angular trials illustrates that such peaks were prominently localized over the frontocentral cortical area. Similarly, we observed that the peaks of ERPs in correct trials were also localized over the central area of the brain. Although both correct and erroneous trials exhibited sequential peaks, the peak-to-peak amplitude was higher

\footnote{Onset of correct angular trials is the same as for erroneous trials; i.e. the moment when the cursor passes over the center.}
in erroneous trials than in correct trials. Significant differences between correct and erroneous amplitudes were observed around 0.35, 0.45 and 0.6 s. (Wilcoxon signed-rank tests, Benjamini and Hochberg’s false discovery rate adjustment, $p < 0.05$). Furthermore, we performed cross-correlation analysis...
Figure 3(b) represents the grand averaged signals at FCz of the distance condition. The ERPs of correct trials exhibited a positive peak at 0.27 s and a negative peak at 0.51 s. The ErrPs exhibited a sequential negative, positive and negative peaks at 0.27 s, 0.37 s and 0.60 s, respectively. Significant differences were observed around 0.30 s and 0.50 s of distance trials. Similar to the angular condition, topographical representation of erroneous angular trials illustrates that those peaks were also mainly localized over the frontocentral cortex, however localization of the peaks were not as clear as the angular conditions. ERPs of correct trials were not localized over the central area of the brain clearly, especially for the positive deflection at 0.31 s. Likewise angular condition, the highest correlation coefficients between correct and erroneous trials, 0.49 ± 0.32 (mean ± std) was observed when the erroneous trials were shifted 0.08 s earlier.

Figure 3(c) depicts the comparison of ErrPs between the erroneous angular and the erroneous distance trials. Noticeably, the first positive peak of the distance condition had a latency of 0.05 s with respect to the angular condition. Moreover, such delay was preserved until the appearance of the negative peaks of the erroneous distance and angular conditions. Significant differences appeared around 0.3 and 0.45 s. We performed statistical analysis on the latency of each positive and negative peak, independently. Latency of the positive peak of erroneous angular and erroneous distance trials were 0.35 ± 0.04 s (mean ± std) and 0.42 ± 0.11 s, respectively (p = 0.007, Wilcoxon’s signed-rank test). Similarly, latency of the negative peak of each erroneous condition also exhibited a significant difference (Erroneous angular trials: 0.59 ± 0.04 s (mean ± std), Erroneous distance trials: 0.66 ± 0.13 s, p = 0.014, Wilcoxon’s signed-rank test). Additionally, we performed a cross-correlation analysis to quantify the temporal delay between the two types of ErrPs. The highest correlation coefficients, 0.65 ± 0.09 (mean ± std), were achieved when the erroneous angular trials were shifted by 0.04 s. Similar to erroneous trials, we compared correct angular and correct distance trials. However, no significant difference was found. Supplementary figures 2 and 3 represent the group and phase delays imposed by the causal filter and the grand-averaged signals with non-causal filtering in the same configuration as figure 3, respectively.

3.3. Synchronous classification
Figure 4 displays the AUC scores for each participant, plus the average over all of them, for all possible pairs of training and test datasets. Table 1 reports the averaged AUC, TNR and TPR plus significance threshold of each measure for each pair. The classification performance outperformed the chance levels in all cases. Similarly, both TNR and TPR outperformed the significance threshold for all pairs of combinations.

In order to assess statistical differences of the classification performances across all pairs of training and test datasets, we performed a two-way repeated measures ANOVA. The first factor was Error Type (i.e. angle or distance), while the second factor was Transfer Learning (i.e. training and test data were from the same condition, combined data was used for training, training and test data were not from the same condition). Specifically, six AUC scores,
Table 1. AUC, TNR, TPR and their significance thresholds of the synchronous classification analysis for each pair of training and test datasets (mean ± std). Decision threshold and the averaged decision threshold during the permutation tests are also indicated.

<table>
<thead>
<tr>
<th></th>
<th>Angle-Angle</th>
<th>Combined-Angle</th>
<th>Distance-Angle</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUC</strong></td>
<td>0.88 ± 0.05</td>
<td>0.87 ± 0.06</td>
<td>0.82 ± 0.04</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.59 ± 0.02</td>
<td>0.59 ± 0.03</td>
<td>0.55 ± 0.02</td>
</tr>
<tr>
<td><strong>TNR</strong></td>
<td>0.85 ± 0.06</td>
<td>0.87 ± 0.03</td>
<td>0.85 ± 0.06</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.77 ± 0.04</td>
<td>0.78 ± 0.04</td>
<td>0.76 ± 0.07</td>
</tr>
<tr>
<td><strong>TPR</strong></td>
<td>0.79 ± 0.10</td>
<td>0.75 ± 0.13</td>
<td>0.68 ± 0.14</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.64 ± 0.03</td>
<td>0.64 ± 0.03</td>
<td>0.64 ± 0.07</td>
</tr>
<tr>
<td>Decision threshold</td>
<td>0.48 ± 0.09</td>
<td>0.51 ± 0.11</td>
<td>0.46 ± 0.14</td>
</tr>
<tr>
<td>Avg. (Permutation test)</td>
<td>0.46 ± 0.03</td>
<td>0.50 ± 0.04</td>
<td>0.47 ± 0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Distance-Distance</th>
<th>Combined-Distance</th>
<th>Angle-Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUC</strong></td>
<td>0.94 ± 0.04</td>
<td>0.94 ± 0.04</td>
<td>0.88 ± 0.07</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.60 ± 0.03</td>
<td>0.61 ± 0.03</td>
<td>0.58 ± 0.02</td>
</tr>
<tr>
<td><strong>TNR</strong></td>
<td>0.89 ± 0.06</td>
<td>0.89 ± 0.06</td>
<td>0.86 ± 0.08</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.72 ± 0.06</td>
<td>0.72 ± 0.07</td>
<td>0.68 ± 0.15</td>
</tr>
<tr>
<td><strong>TPR</strong></td>
<td>0.89 ± 0.09</td>
<td>0.87 ± 0.06</td>
<td>0.81 ± 0.11</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.67 ± 0.06</td>
<td>0.68 ± 0.06</td>
<td>0.70 ± 0.10</td>
</tr>
<tr>
<td>Decision threshold</td>
<td>0.44 ± 0.06</td>
<td>0.49 ± 0.07</td>
<td>0.46 ± 0.13</td>
</tr>
<tr>
<td>Avg. (Permutation test)</td>
<td>0.39 ± 0.07</td>
<td>0.44 ± 0.06</td>
<td>0.38 ± 0.12</td>
</tr>
</tbody>
</table>

TNR and TPR were computed for a single participant, one value per condition. When the training and the test data were consistent or the combined data was used for training, averaged value across test folds were computed. In the inconsistent cases, the computed performance was directly used for the statistical analysis. This process resulted in the $11 \times 6$ matrix (number of participants × number of conditions) for AUC, TNR and TPR. Two-way repeated measures ANOVA on AUC scores showed no interaction effect of the factors Error Type × Transfer Learning ($F(2,20) = 0.64, p = 0.54$). However, we observed a significant effect of Error Type ($F(1,10) = 17.4, p = 0.002$) and Transfer Learning ($F(2,20) = 22.1, p < 0.001$). The subsequent post-hoc analysis revealed significant differences between Angle-Angle against Distance-Distance and against Distance-Angle, Combined-Angle against Combined-Distance and against Distance-Angle, Distance-Angle against Angle-Distance, Distance-Distance against Angle-Distance, and Combined-Distance against Angle-Distance as shown in figure 4.

Similarly, two-way repeated ANOVA on TNR and TPR also exhibited no significant interaction between the factors (TNR: $F(2,20) = 0.58, p = 0.57$, TPR: $F(2,20) = 0.32, p = 0.73$), but showed significant main effect of Transfer Learning for both TNR and TPR (TNR: $F(2,20) = 5.93, p = 0.009$, TPR: $F(2,20) = 15.4, p < 0.01$) and Error Type only for TPR (TNR: $F(1,10) = 1.96, p = 0.19$, TPR: $F(1,10) = 14.4, p = 0.004$). The post-hoc analysis revealed significant differences between Combined-Distance against Angle-Distance for both TNR and TPR, Combined-Angle against Distance-Angle, Angle-Angle against Distance-Angle, Angle-Distance against Angle-Distance only for TPR. Table 2 reports the results of the post-hoc analysis.

### 3.4. Asynchronous classification

Figure 5(a) represents the averaged continuous posterior probabilities between (-0.5, 1.5) s with respect to the onset of angular events for correct trials generated by the three types of decoders. Noticeably, the three types of decoders yielded the highest posterior probability at 0.98 s (angle decoder: $0.70 ± 0.02$, combined decoder: $0.71 ± 0.02$, and distance decoder: $0.62 ± 0.02$).
Table 2. Results of the post-hoc analysis for AUC of synchronous classification. Estimated difference indicates the estimated difference between the corresponding two marginal means. Standard error represents the standard error of the estimated difference.

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Estimated difference</th>
<th>Standard error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle-Angle</td>
<td>Distance-Distance</td>
<td>−0.059</td>
<td>0.013</td>
<td>0.001</td>
</tr>
<tr>
<td>Angle-Angle</td>
<td>Combined-Angle</td>
<td>0.014</td>
<td>0.006</td>
<td>0.163</td>
</tr>
<tr>
<td>Angle-Angle</td>
<td>Distance-Angle</td>
<td>0.059</td>
<td>0.011</td>
<td>0.001</td>
</tr>
<tr>
<td>Combined-Angle</td>
<td>Combined-Distance</td>
<td>−0.066</td>
<td>0.015</td>
<td>0.001</td>
</tr>
<tr>
<td>Combined-Angle</td>
<td>Distance-Angle</td>
<td>0.046</td>
<td>0.009</td>
<td>0.002</td>
</tr>
<tr>
<td>Distance-Angle</td>
<td>Angle-Distance</td>
<td>−0.056</td>
<td>0.018</td>
<td>0.010</td>
</tr>
<tr>
<td>Distance-Distance</td>
<td>Combined-Distance</td>
<td>0.006</td>
<td>0.006</td>
<td>0.650</td>
</tr>
<tr>
<td>Distance-Distance</td>
<td>Angle-Distance</td>
<td>0.062</td>
<td>0.016</td>
<td>0.009</td>
</tr>
<tr>
<td>Combined-Distance</td>
<td>Angle-Distance</td>
<td>0.057</td>
<td>0.013</td>
<td>0.005</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Estimated difference</th>
<th>Standard error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle-Angle</td>
<td>Distance-Distance</td>
<td>−0.040</td>
<td>0.017</td>
<td>0.047</td>
</tr>
<tr>
<td>Angle-Angle</td>
<td>Combined-Angle</td>
<td>−0.016</td>
<td>0.018</td>
<td>0.790</td>
</tr>
<tr>
<td>Angle-Angle</td>
<td>Distance-Angle</td>
<td>0.002</td>
<td>0.024</td>
<td>0.999</td>
</tr>
<tr>
<td>Combined-Angle</td>
<td>Combined-Distance</td>
<td>−0.023</td>
<td>0.021</td>
<td>0.293</td>
</tr>
<tr>
<td>Distance-Angle</td>
<td>Distance-Angle</td>
<td>0.018</td>
<td>0.017</td>
<td>0.653</td>
</tr>
<tr>
<td>Distance-Distance</td>
<td>Combined-Distance</td>
<td>−0.007</td>
<td>0.031</td>
<td>0.814</td>
</tr>
<tr>
<td>Distance-Distance</td>
<td>Angle-Distance</td>
<td>0.0003</td>
<td>0.010</td>
<td>0.999</td>
</tr>
<tr>
<td>Distance-Distance</td>
<td>Angle-Distance</td>
<td>0.034</td>
<td>0.018</td>
<td>0.216</td>
</tr>
<tr>
<td>Combined-Distance</td>
<td>Angle-Distance</td>
<td>0.034</td>
<td>0.010</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Figures 5(c) and (d) illustrate the averaged continuous posterior probability with respect to the onset of distance events for correct and erroneous trials, respectively. Consistently with figures 5(a) and (c) exhibited the highest posterior probability for correct distance trials was reached at 0.80 s for correct trials (distance decoder: 0.37 ± 0.03 (mean ± std), combined decoder: 0.36 ± 0.02, and angle decoder: 0.30 ± 0.02). Similar to figures 5(b) and (d) shows the highest posterior probability at 0.99 s with respect to the onset of distance events for erroneous trials (distance decoder: 0.75 ± 0.02 (mean ± std), combined decoder: 0.76 ± 0.02, and angle decoder: 0.66 ± 0.03).

Figure 6(a) illustrates the averaged asynchronous classification performance using the balanced classification accuracy (BAC), i.e. the average of TNR and TPR. Table 3 reports the averaged asynchronous BAC, TNR, TPR, and their corresponding significance thresholds together with their standard deviation, for all possible pairs of training and test conditions. Noticeably, comparable performances were obtained regardless of training data, especially between consistent and combined conditions. We observed that our approach outperformed the significance threshold for BAC and TPR for all conditions, but not for TNR. Similar to the synchronous classification analysis, we computed a two-way repeated measures ANOVA on BAC. We found no interaction effect of the factors Error Type × Transfer Learning (F(2,20) = 1.25, p = 0.31), nor a significant effect of Error Type (F(1,10) = 1.43, p = 0.26). However, we observed a significant effect of Transfer Learning (F(2,20) = 16.49, p < 0.001). We performed additional two two-way repeated measures ANOVAs on TNR and TPR, independently. Similar to BAC, we found no significant interaction between the factors (TNR: F(2,20) = 1.19, p = 0.32, TPR: F(2,20) = 0.46, p = 0.64), nor significant main effects of Error Type (TNR: F(1,10) = 0.01, p = 0.92,
TPR: $F(1, 10) = 1.40, p = 0.26)$. However, we observed a significant main effect of Transfer Learning only on TPR, but not on TNR (TNR: $F(2, 20) = 0.66, p = 0.53$, TPR: $F(2, 20) = 19.1, p < 0.001$). Thus, the subsequent post-hoc analysis was performed on BAC and TPR. We observed a significant difference in BAC and TPR for the pairs of Combined-Angle against Distance-Angle and Combined-Distance against Angle-Distance. The pair of Distance-Distance against
Figure 6. Asynchronous classification accuracy for different pairs of training and test datasets. (a) Averaged balanced classification accuracy (BAC) and standard deviation are represented by bar and error-bar, respectively. Balanced classification accuracy (i.e., average of TNR and TPR) of each participant is depicted by filled dots for each pair of training and test data. X axis represents the pair of ‘training data—test data’. The averaged significance threshold (\(\alpha = 0.05\)) is shown by the black lines for each condition.

(b)–(d) TNR and TPR at each threshold \(\tau\) for the two types of trials averaged over participants and standard deviation for angular, combined and distance training conditions. Red line represents TPR for angular trials. Blue line represents the TPR for distance trials, while black lines represent TNR for both kind of trials. The significance threshold of TNR and TPR is shown by the dashed line of the corresponding color.

Angle-Distance was found to be significantly different only in BAC.

Figures 6(b)–(d) show the TNR and TPR for each decision threshold for all the three training conditions. Noticeably, TNR increased along with the threshold in all conditions; however, TPR had a bell-shaped behavior with positive peak of \(0.61 \pm 0.17\) (mean \(\pm\) std) at a threshold of \(0.73\) for angular condition, \(0.63 \pm 0.16\) at 0.78 for combined condition and \(0.65 \pm 0.22\) at 0.76 for distance condition. Note
Table 3. Balanced classification accuracy (BAC), TNR, TPR, and their significance thresholds of asynchronous classification for each pair of training and test datasets (mean ± std). Decision threshold and the averaged decision threshold during the permutation tests are also indicated.

<table>
<thead>
<tr>
<th></th>
<th>Angle-Angle</th>
<th>Combined-Angle</th>
<th>Distance-Angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC</td>
<td>0.65 ± 0.09</td>
<td>0.66 ± 0.10</td>
<td>0.62 ± 0.10</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.50 ± 0.002</td>
<td>0.50 ± 0.001</td>
<td>0.50 ± 0.003</td>
</tr>
<tr>
<td>TNR</td>
<td>0.80 ± 0.06</td>
<td>0.81 ± 0.04</td>
<td>0.84 ± 0.09</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.98 ± 0.02</td>
<td>0.99 ± 0.002</td>
<td>0.99 ± 0.01</td>
</tr>
<tr>
<td>TPR</td>
<td>0.50 ± 0.20</td>
<td>0.50 ± 0.20</td>
<td>0.40 ± 0.26</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.12 ± 0.06</td>
<td>0.09 ± 0.04</td>
<td>0.10 ± 0.04</td>
</tr>
<tr>
<td>Decision threshold</td>
<td>0.84 ± 0.07</td>
<td>0.84 ± 0.06</td>
<td>0.82 ± 0.11</td>
</tr>
<tr>
<td>Avrg. (Permutation test)</td>
<td>0.98 ± 0.03</td>
<td>0.99 ± 0.003</td>
<td>0.98 ± 0.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Distance-Distance</th>
<th>Combined-Distance</th>
<th>Angle-Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC</td>
<td>0.69 ± 0.12</td>
<td>0.69 ± 0.13</td>
<td>0.65 ± 0.12</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.50 ± 0.003</td>
<td>0.50 ± 0.001</td>
<td>0.50 ± 0.002</td>
</tr>
<tr>
<td>TNR</td>
<td>0.84 ± 0.05</td>
<td>0.80 ± 0.07</td>
<td>0.83 ± 0.12</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.98 ± 0.01</td>
<td>0.99 ± 0.002</td>
<td>0.99 ± 0.02</td>
</tr>
<tr>
<td>TPR</td>
<td>0.54 ± 0.26</td>
<td>0.58 ± 0.20</td>
<td>0.43 ± 0.21</td>
</tr>
<tr>
<td>Significance threshold</td>
<td>0.14 ± 0.06</td>
<td>0.09 ± 0.04</td>
<td>0.11 ± 0.04</td>
</tr>
<tr>
<td>Decision threshold</td>
<td>0.82 ± 0.10</td>
<td>0.84 ± 0.06</td>
<td>0.85 ± 0.09</td>
</tr>
<tr>
<td>Avrg. (Permutation test)</td>
<td>0.98 ± 0.01</td>
<td>0.99 ± 0.003</td>
<td>0.99 ± 0.01</td>
</tr>
</tbody>
</table>

that when the significance threshold is too low (e.g. 0.1), all trials are classified as error trials, because detection of error happens earlier than the onset of the erroneous events, what yields a low TPR. Similar to the previous studies [45, 56] which performed tuning of the decision threshold based on the product of TPR and TNR, we used BAC which also takes into account both TPR and TNR. Choosing the decision threshold based on BAC led to an improvement of ~10% for BAC compared to the case in which the decision threshold was tuned solely on TPR.

As shown in figures 6(b)–(d), we computed the significance threshold (α = 0.05) of TNR and TPR for each possible threshold, summarized in table 3. The maximum significance threshold of TPR was 0.32 ± 0.02 (mean ± std) at a threshold of 0.91 for angular condition, 0.31 ± 0.01 at 0.92 for combined condition and 0.35 ± 0.02 at 0.89 for distance condition. The significance threshold of TNR remained below 0.1 until 0.8 of decision threshold, and exponentially increased to 1.0 between (0.8, 1.0) of threshold for all the conditions. Given the non-bell-shaped trend of the TNR, and corresponding significance threshold, for all kinds of trials, the significance threshold for BAC was tuned to maximize TNR, which resulted in the significance threshold for BAC to be close to 0.5.

4. Discussion

In the present study, the experimental continuous feedback protocol was designed to induce ErrPs in two different conditions, i.e. at predicted or unpredictable states. The former condition is the case in which users can infer where and when an erroneous action may occur, i.e. angular event; while for the latter condition users cannot precisely anticipate eventual errors, i.e. distance event. We used 16 EEG channels around the front-central area of the brain to capture the maximum modulation of ErrPs [57]. The choice of the number and location of EEG electrodes aimed at recovering activity of the anterior cingulate cortex (ACC), which previous studies [35, 36] demonstrated to be the main source of ErrPs.

Furthermore, other studies [26, 34] using the same EEG placement observed clear elicitation of ErrPs and achieved similar classification performance to other studies with a larger number of electrodes. Although previous EEG studies [19, 58] reported the simultaneous activation of somatosensory cortex (Brodmann areas 5 and 7) upon the perception of the errors, these studies reported topographical representations with patterns of activation on the central area, which is covered by our 16-channel setup.

As shown in figure 3, and in line with the previous studies [11, 19–24, 27, 29, 34–37, 50], ErrPs
were composed of similar sequential peaks in both conditions. However, ErrPs associated to the distance condition had a significantly higher latency than ErrPs elicited by the angular condition, confirming our hypothesis that users need to deploy more cognitive effort for detecting distance errors. The relationship between the latency of ErrPs and cognitive workload have been previously shown for discrete actions [60, 61], while Pe reflects conscious perception of errors [62]. However, participants did not commit erroneous actions in our experiment, but their advantage [49, 50, 63]. In particular, the methods CCA and xDAWN have been reported to yield systematic increases in classification performance of time-locked ERPs [50, 63], with a slightly superiority of CCA. We have, therefore, adopted the CCA spatial filter for our analysis, which has demonstrated to perform well not only during synchronous classification, as expected, but also during asynchronous detection of ErrPs. In fact, replacing CCA with a common average reference (CAR), degrades asynchronous decoding significantly by 10% or more for every condition as expected, but also during asynchronous detection of ErrPs. In fact, replacing CCA with a common average reference (CAR), degrades asynchronous decoding significantly by 10% or more for every condition reported in table 3 (data not shown).

Table 4. Results of the post-hoc analysis for AUC of synchronous classification. Estimated difference indicates the estimated difference between the corresponding two marginal means. Standard error represents the standard error of the estimated difference.

<table>
<thead>
<tr>
<th>BAC</th>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Estimated difference</th>
<th>Standard error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle-Angle</td>
<td>Distance-Distance</td>
<td>−0.037</td>
<td>0.025</td>
<td>0.179</td>
<td></td>
</tr>
<tr>
<td>Angle-Angle</td>
<td>Combined-Angle</td>
<td>−0.006</td>
<td>0.014</td>
<td>0.961</td>
<td></td>
</tr>
<tr>
<td>Angle-Angle</td>
<td>Distance-Angle</td>
<td>0.035</td>
<td>0.018</td>
<td>0.236</td>
<td></td>
</tr>
<tr>
<td>Combined-Angle</td>
<td>Combined-Distance</td>
<td>−0.033</td>
<td>0.023</td>
<td>0.177</td>
<td></td>
</tr>
<tr>
<td>Combined-Angle</td>
<td>Distance-Angle</td>
<td>0.041</td>
<td>0.009</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Distance-Angle</td>
<td>Angle-Distance</td>
<td>−0.008</td>
<td>0.025</td>
<td>0.747</td>
<td></td>
</tr>
<tr>
<td>Distance-Distance</td>
<td>Combined-Distance</td>
<td>−0.003</td>
<td>0.011</td>
<td>0.993</td>
<td></td>
</tr>
<tr>
<td>Distance-Distance</td>
<td>Angle-Distance</td>
<td>0.063</td>
<td>0.016</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>Combined-Distance</td>
<td>Angle-Distance</td>
<td>0.066</td>
<td>0.015</td>
<td>0.005</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TPR</th>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Estimated difference</th>
<th>Standard error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle-Angle</td>
<td>Distance-Distance</td>
<td>−0.043</td>
<td>0.055</td>
<td>0.452</td>
<td></td>
</tr>
<tr>
<td>Angle-Angle</td>
<td>Combined-Angle</td>
<td>−0.008</td>
<td>0.018</td>
<td>0.967</td>
<td></td>
</tr>
<tr>
<td>Angle-Angle</td>
<td>Distance-Angle</td>
<td>0.100</td>
<td>0.395</td>
<td>0.086</td>
<td></td>
</tr>
<tr>
<td>Combined-Angle</td>
<td>Combined-Distance</td>
<td>−0.078</td>
<td>0.046</td>
<td>0.116</td>
<td></td>
</tr>
<tr>
<td>Combined-Angle</td>
<td>Distance-Angle</td>
<td>0.108</td>
<td>0.031</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>Distance-Angle</td>
<td>Angle-Distance</td>
<td>−0.028</td>
<td>0.056</td>
<td>0.630</td>
<td></td>
</tr>
<tr>
<td>Distance-Distance</td>
<td>Combined-Distance</td>
<td>−0.043</td>
<td>0.035</td>
<td>0.571</td>
<td></td>
</tr>
<tr>
<td>Distance-Distance</td>
<td>Angle-Distance</td>
<td>0.115</td>
<td>0.046</td>
<td>0.095</td>
<td></td>
</tr>
<tr>
<td>Combined-Distance</td>
<td>Angle-Distance</td>
<td>0.158</td>
<td>0.036</td>
<td>0.004</td>
<td></td>
</tr>
</tbody>
</table>
achieved performances similar to the state-of-the-art [35, 51, 64, 65], with a BAC around 65% for both predicted (i.e. angular) and unpredicted (i.e. distance) states. Our results indicated slightly lower classification performance compared to the previous study that performed asynchronous classification [51,56]. This is primarily due to the presence of event-related potentials in the correct trials, which led to an increase of the posterior probability in the correct trials at 0.89 s for correct angular trials and 0.80 s for correct distance trials, respectively. Such increase of posterior probability made the asynchronous classification harder than the previous studies in which event-related potentials were not present in the correct trials. Nevertheless, the most relevant finding is that, contrarily to suggestions by previous studies, ErrP decoders trained in a given condition successfully transferred to the other condition, with the slight loss of classification performance. This was the case for both, synchronous as well as asynchronous ErrP decoding. Specifically, although for both synchronous and asynchronous decoding applying the classifier trained in one of the conditions (angular or distance) to the other condition yielded a significant decrease in performance, except for the case in which transferring the angular decoder to the distance trials during asynchronous classification, this decrease was small compared to the case in which training and test conditions are consistent (synchronous decoding: ∼6%, table 1, table 2 and figure 4: from 88% to 82% in the angular condition; from 94% to 88% in the distance condition; asynchronous decoding: ∼6%, table 3, table 4 and figure 4: from 69% to 63% in the distance condition) and still well above the significance threshold of classification performance. This was the case for both, synchronous as well as asynchronous ErrP decoding. Furthermore, we observed that a single decoder trained by mixture of the two types of events yields similar classification performance compared to the case where training and test data are consistent. This novel result enlarges the potential use of ErrPs during natural operation of brain-controlled devices as it will not be necessary to have different decoders for each kind of erroneous conditions.

Future work will need to address a few open issues, such as how to improve asynchronous ErrP decoding performance to the level of synchronous decoding, whether results achieved in the reported offline, pseudo-online analysis are confirmed during actual online experiments, i.e. does the decoder trained with both types of ErrPs can be successfully applied to ErrPs of both types of predicted and unpredicted states without loss of classification performance, and whether other continuous feedback tasks—especially involving physical devices—also exhibit the same relationship between ErrPs for predicted and unpredicted events.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgment

This work was supported by the Hasler Foundation, Switzerland.

ORCID iDs

Fumiaki Iwane  
https://orcid.org/0000-0002-9659-4127
References


[31] Hill H 2009 An event-related potential evoked by movement planning is modulated by performance and learning in visuomotor control Exp. Brain Res. 195 519–29


[34] Iturrate I, Chavarriaga R, Montesano L, Mengue J and Millán J D R 2014 Latency correction of event-related potentials between different experimental protocols J. Neural. Eng. 11 036005


[40] Yazmir B and Reiner M 2017 Monitoring brain potentials to guide neurorehabilitation of tracking impairments 2017 Int. Conf. on Rehabilitation Robotics (ICORR) pp 983–8


[47] Grandchamp R and Deforme A 2011 Single-trial normalization for event-related spectral decomposition reduces sensitivity to noisy trials Frontiers Psychol. 2 236


[51] Spüler M 2017 Spatial filtering of EEG as a regression problem 7th Graz Brain–Computer Conf

[52] Spüler M 2018 Masked and unmasked error-related potentials during continuous control and feedback J. Neural Eng. 15 036031


Case Report

The AI-Atlas: Didactics for Teaching AI and Machine Learning On-Site, Online, and Hybrid

Thilo Stadelmann 1,*,†, Julian Keuzenkamp 2, Helmut Grabner 3 and Christoph Würsch 4


Abstract: We present the “AI-Atlas” didactic concept as a coherent set of best practices for teaching Artificial Intelligence (AI) and Machine Learning (ML) to a technical audience in tertiary education, and report on its implementation and evaluation within a design-based research framework and two actual courses: an introduction to AI within the final year of an undergraduate computer science program, as well as an introduction to ML within an interdisciplinary graduate program in engineering. The concept was developed in reaction to the recent AI surge and corresponding demand for foundational teaching on the subject to a broad and diverse audience, with on-site teaching of small classes in mind and designed to build on the specific strengths in motivational public speaking of the lecturers. The research question and focus of our evaluation is to what extent the concept serves this purpose, specifically taking into account the necessary but unforeseen transfer to ongoing hybrid and fully online teaching since March 2020 due to the COVID-19 pandemic. Our contribution is two-fold: besides (i) presenting a general didactic concept for tertiary engineering education in AI and ML, ready for adoption, we (ii) draw conclusions from the comparison of qualitative student evaluations (n = 24–30) and quantitative exam results (n = 62–113) of two full semesters under pandemic conditions with the result of previous years (participants from Zurich, Switzerland). This yields specific recommendations for the adoption of any technical curriculum under flexible teaching conditions—be it on-site, hybrid, or online.

Keywords: flexible educational design; e-learning; constructivism; design-based research; COVID-19; post-pandemic tertiary engineering education; artificial intelligence

1. Analysis and Exploration

1.1. The Problem of Teaching Artificial Intelligence as a Foundational Subject

Students enter their careers in a time that awaits nothing short of a digital disruption [1]. The core of the disruptive potential of the digital transformation is provided by technological developments, foremost by Artificial Intelligence (AI) and its currently most prominent branch, Machine Learning (ML) [2]. While technical in nature, AI and ML have the potential to disrupt society on all levels, including business, public service, justice, art, science and health [3]. Numerous scientific articles are published daily [4], and coverage in popular media contributes to these technological developments not just a scientific, but also a mainstream hype [5]. The hype leads to misconceptions about what the technology is capable of, or designed to do, including that it could be concerned with human intelligence or creating conscious machines [6] (while the reality boils down to
mere “complex problem solving” [7]). This exemplary problem highlights the necessity of a foundational understanding of AI for engineers in two respects: (i) the understanding needs to be solid, i.e., rooted in the foundations of the discipline rather than purely in its applications, tools or business value, to manage expectations; (ii) like algebra and other foundational subjects, the understanding of AI needs to serve as an underlying methodological framework for an increasing number of future engineering efforts, i.e., it needs to serve as a foundation for future careers in engineering rather than merely as an interesting fashionable specialization.

We think of AI and its sub-discipline ML as one of the five pillars of the discipline of computer science (besides theoretical, practical, technical, and applied computer science [8]). As such, it requires similar treatment to other foundational subjects (e.g., thorough establishment of basic ideas and theory), but holds a peculiarity: different from foundations, such as algebra, AI and ML already build upon a body of knowledge from computer science. This implies a later slot in respective programs, and hence more experienced students who can better judge the impact of AI on other aspects of their profession or society as a whole. However, teaching AI as a foundational aspect of a technical education was until recently largely neglected in curricula according to surveys [9], and novel data science courses [10] do not cover the same terrain.

We, as educators and researchers in this context, thus, are faced with the problem of having to prepare experienced students with diverse professional backgrounds for the reality of AI and ML applications and their implications. For this, we need university courses that provide a solid theoretical foundation while providing opportunities and a mind-set for applied life-long-learning, especially in tertiary engineering education at universities of applied science. We need to teach solid AI and ML foundations to this diverse, technical audience while the field is rapidly evolving, students may have initial misconceptions, and the outcome has to be practice-oriented and relevant for a broad range of careers. In short: we need a didactic concept to solve for this problem by providing a powerful heuristic—a set of coherent best practices—for curriculum and didactic planning and implementation. While many courses exist, and new ones are rapidly created to keep up with the current surge, to our knowledge, no such didactic concept exists explicitly. Earlier AI didactics, like References [11,12], focused only on special teaching situations or one of the sub-discipline of AI.

In this paper, we report on the iterative design of two courses based on the “AI-Atlas” didactic concept. We discuss how the concept itself was developed in specific response to the aforementioned challenges induced by the current surge of AI, and how the resulting course design was adapted based on student feedback and the rapidly changing context of COVID-19. Specifically, our case report is based on the structure of Design-Based Research (DBR) [13] to present the AI-Atlas (didactic concept) and two exemplary implementations (concrete courses), and evaluate the merits of the concept based on student results and feedback. The courses, “Artificial Intelligence I” (referred to as the “AI course” below) for undergraduate students of computer science, and “Machine Learning” (“ML course”) within an interdisciplinary graduate program in engineering, are described in detail in Appendix A. The atlas (or map) metaphor itself that gave rise to the AI-Atlas concept and its underlying design choices is introduced in Section 2.1. We chose to present our findings as a case-report rather than a research article because there are certain limitations regarding our evaluation data as explained in Section 3.2. Our intention is, therefore, to introduce our didactic concept and to motivate educators to test it in order to gather more experience and evidence for the suitability of the AI-Atlas as a flexible basis for designing a curriculum for AI and ML courses in tertiary education.

1.2. Existing AI Curricula and Their Relation to Our Didactic Concept

It is beyond the scope of this case report to present a comprehensive survey of related literature on AI and ML teaching methodology. Nevertheless, the AI-Atlas fits well within the recent discussion on AI curricula: like Williams et al. [14], constructionist ideas are
central to our didactic design as discussed in Section 2.3. Moreover, the origins of the AI-Atlas [15,16] evolved from teaching best practices of the involved lecturers, putting this case report in line with similar post-hoc analyses, for example, References [17–19].

Chiu and Chai [20] reflect AI curriculum development for K-12 [21] by teachers with and without prior exposure to AI training. They use the psychological framework of Self-Determination Theory (SDT) to understand how teachers’ motivation controls curriculum design. Further, the authors use the four planning paradigms of curriculum as either content, product, process, or practice to distinguish fundamentally different ways to think about course development and respective didactic designs. While SDT can also be used to explain our motivation for the specific design of the AI-Atlas, our concept can be classified as being a blend of content (the syllabus is important to us), product (we care about achieved educational objective), and practice (we frequently connect the learning experience to real-world applications); see Section 2.5.

Finally, Li et al. [22] explore the fit between university AI curricula and the demands of the job market. Our design with its focus on professional applicability can be seen as a solution to the problems they identify within their analyzed courses: What distinguishes our concept from the many other adoptions of, e.g., the AIMA textbook (cf. Reference [23]), is an end-to-end focus on the usefulness of the taught foundations in daily application that should lead to successful transfer into personal problem-solving skills and professional practice. This implies to be less focused on covering a large amount of different AI/ML algorithms, but rather teach the underlying relationships. Of course, practical relevance is a major concern for any course at a university of applied sciences. While we cannot compare the AI-Atlas to a previous AI course at our institutions (it was developed for the initial design of the first AI and ML courses here), we notice that our approach takes an atypical route to this destination: most courses are made more “applied” by stripping them of theoretical concepts. However, we put theoretical foundations at the center. The reasoning for this seemingly uncommon choice follows below.

2. Design and Construction

In this section, we explain why we named our didactic concept the AI-Atlas and what the core mentality is it means to convey. We also look at three didactic principles that guided us in its design and three aspects about AI and ML teaching we believe it needs to cover in order to solve the above stated problem before breaking them down into the suggestion of specific didactic settings. Together, the principles and ideas layed out in the following five subsections constitute the AI-Atlas didactic concept.

2.1. The Atlas Metaphor

In the late 16th century, Gerhard Mercator’s “Atlas Sive Cosmographicae Meditationes de Fabrica Mundi et Fabricati Figura” combined maps and associated explanations of the known world [24]. They were used by generations to explore, push boundaries, and further trade and development [25]. For an atoll, for example, they would show all individual islands with their border, list their characteristics, and show their relation to each other. This aids travelers by allowing them to plan the most effective or efficient route based on their current needs or interest. However, it does not set an pre-conceived path and, if a new island is discovered, it can be added to the atlas without disrupting existing knowledge.

Within the AI-Atlas, we think of the sub-disciplines of AI—such as, e.g., search, planning, and machine learning—as individual islands of an atoll: well developed in themselves and somewhat related to each other, but missing a direct connection. Hence, to solve the problem we have identified, future generations of professionals need the analogue to what Mercator gave to his contemporaries: an atlas to the world of AI (cf. Figure 1). An explorer can profit from the help of this atlas to get an overview and find the best path given a specific journey (i.e., application). It highlights main routes (i.e., baseline approaches), special landmarks (e.g., important algorithms, killer applications), and borders (i.e., limits
of the approaches, future work) but never restricts learners to knowing or using only a single path.

![Image](image_url)

**Figure 1.** Illustration of the AI-Atlas metaphor: the sub-disciplines of AI are like an atoll that is best navigated with the help of an atlas. The AI-Atlas didactic concept provides the means, to be employed by educators, to let such atlases emerge in their learner’s minds. Image credit: Colorbox, used with permission.

Our AI-Atlas didactic concept contains the means to let the actual atlas emerge only in a learner’s mind. It is, thus, created by analog means and stored in analog form (in natural neural networks) and is not manifested in some digital format on the learner’s computer. This aspect of the metaphor underlines the AI philosophy underlying the didactic concept and derived courses: artificial intelligence is not primarily replacing human intelligence, and machine learning does not render human learning unnecessary, just like digital does not primarily replace analog, but augments it [26]. AI, thus, finds an optimal environment for application where human and machine complement each other with their respective strengths and weaknesses.

2.2. The Core Mentality: AI Professionals Are Explorers

The discipline of AI and its major tool of ML do not have a single goal (“creating intelligence”), but rather offer a methodological toolbox to approach multiple targets (“solving complex problems”) [7]. Thus, at their core, they are not constituted by technology as much as by a specific mentality: since AI’s inception as a discipline in the 1950s, AI researchers notoriously approached the kind of problems with creativity and pragmatism that had been laid aside by fellow researchers from other disciplines as “too hard” [27]. In other words, AI researchers explored previously unknown territories. They did so by employing an interdisciplinary “let’s do it” mentality. Today, this mentality distinguishes the work of the AI professional from other modeling approaches used by software engineers, database designers or statisticians, although skills in all these areas are relevant for success in and with AI or ML. The AI-Atlas not only acknowledges but actively hones this explorer mentality [28]. It does so by building on a set of corresponding didactic principles.

2.3. Didactic Principles

2.3.1. Principle of “Doing It Yourself”

Since the late 1980s, constructivist ideas have increasingly found their way into pedagogy, as well as into discussions on the design of teaching-learning environments. Constructivism postulates that individuals do not take over information faithfully from external sources, but actively construct it through social negotiation processes based on previously made experiences. Knowledge is situation-specific and must be actively and independently linked by the individual to prior knowledge [29]. It follows that teaching-
learning settings must be designed in such a way that learners are given the opportunity to actively engage with the learning content, as well as the associated problems and to relate these to their prior experiences, whereby active engagement can be of both visible and non-visible nature.

2.3.2. Principle of “Intrinsic Motivation”

In their Self-Determination Theory of motivation (SDT), Deci and Ryan [30] explain the relationship between motivation, learning, and the influence of the social environment on the fulfillment of basic needs. Intrinsically motivated behavior is associated with individuals freely engaging with the subject matter and striving of their own accord to understand phenomena and master activities that appear to them to be at least personally highly significant. As a further component of their theory, Deci and Ryan postulate three basic human needs that motivate behavior: (i) the need to experience competence, (ii) the need for social inclusion, and (iii) the need for autonomy. Deci and Ryan now assume that the striving for satisfaction of these three basic needs explains why persons pursue certain goals of action and why certain actions are more likely to be perceived as motivating by themselves.

2.3.3. Principle of “EEE” (Good Explanation, Enthusiasm, and Empathy)

According to Winteler [31], the following characteristics of university teaching promote student learning: the instructor’s preparation, organization of the course, clarity and comprehensibility, perceived efficiency of the teaching, the instructor’s openness to questions and discussion, and openness to other opinions. Helmke and Schrader [32] reduce the state of research on key characteristics of successful university teaching to the short formula “EEE”: (i) good explanation, which facilitates information processing and arouses curiosity and factual interest; (ii) commitment and enthusiasm, i.e., the infectious enthusiasm of the lecturer about the content; and (iii) empathy, by which they include personal appreciation of students, openness to problems, and efforts to obtain feedback to better adapt teaching. The fact that the didactic setup of a course is well planned and fine-tuned is a “conditio sine qua non”, i.e., a necessary but not sufficient condition for a successful course. Nevertheless, a committed, authentic (i.e., experienced in the field), and enthusiastic teacher, open for questions and igniting curiosity and factual interest, can bring the majority of the students to engage in the topic and start the learning process in a self-motivated manner.

2.4. The Aspects of Establishing AI Foundations

The above didactic principles need to be combined with the proper mediation of AI and ML foundations, if the AI-Atlas is to successfully guide our explorers-in-training. We suggest that there are three aspects to which the principles need to be applied: canonization, deconstruction, and cross-linkage.

2.4.1. Canonization

The aforementioned hype [5] around AI, and especially deep learning, and the daily growth of scientific literature on the topic [4] make a proper selection of content a key aspect of teaching AI and ML. Hence, a key aspect of the AI-Atlas is to suggest a timely selection of materials that emphasize topics with future relevance alongside their historic development, thereby making the overarching principles that stood the test of time stand out. This is given priority over intriguing detail or formal derivations. Thus, for, e.g., a specific implementation of the AI-Atlas for an introduction to AI course, canonization means that we ensure to teach the full canon of relevant methods (ranging from heuristic search and logical planning to machine learning). We link each of these areas with practice (e.g., controlling a fashionable browser game, building a dragnet investigation system, or decision support for second-hand vehicles). This way, students see for themselves that not only the currently most fashionable methodology, ML, and, within ML not only neural networks, have practical relevance.
2.4.2. Deconstruction

Due to the current extensive media coverage of AI and ML, many misconceptions about the field abound in prospective students (such as the focus of the field being the understanding of human intelligence or the creation of conscious machines [6]). Thus, an important aspect of teaching according to the AI-Atlas is a form of demystification that keeps the original motivation of the students and channels it into more realistic, sustainable paths. We suggest to support students with forming a personal view through critical engagement with scientific texts and programming tasks, which they then present in own write-ups and oral discussions.

2.4.3. Cross-Linkage

Both aspects above—a stable body of knowledge in AI and ML fundamentals and careful treatise of real and misguided excitement—become a firm foundation given the third ingredient: a dense network of cross-references to other subjects in the study program that is compatible with the different occupations of a professional career in engineering and related fields. The AI-Atlas suggests to teach AI not only to future scientists but also to software developers, data scientists, or process engineers, acknowledging the future importance of AI methodology in any field.

2.5. The AI-Atlas in Practice: Suggested Didactic Settings to Combine It All

Building on the principles described above, we suggest to adopt the following specific didactic settings for any AI or ML course facing the problems outlined in Section 1.1. Section 3 then continues to evaluate to which extent employing these means in the two exemplary courses mentioned above achieves these ends. Nevertheless, the following subsections already make frequent references to examples in the AI course and ML course to put the abstract suggestions into concrete terms.

2.5.1. Basic Didactic Setting

As laid out in the theoretical framework above, active engagement is a mandatory key component in an AI-Atlas inspired course. One way to increase engagement is to work with small to medium classes. For example, both exemplary courses had only 30 to 60 students. Courses should, therefore, build upon the “lecture + lab” format widespread in engineering education: weekly lectures accompanied by lab exercises with a roughly even time-split between them. However, we make important adaptations to foster active engagement as follows.

For lectures, the students should read weekly portions of well-established text books as accompaniment to the lectures, e.g., Reference [33] on AI and References [34,35] on ML, complemented where necessary by shorter articles (e.g., References [26,36]). The conveyed anecdotes and historical notes therein specifically contribute to the students’ socialization in the discipline of AI and the field of ML. The lectures themselves connect such context with problem awareness and technical solutions without degenerating into pure 90-min talks that would push learners into passive consuming roles (cf. Section 2.5.5).

Labs on the other hand should go beyond just programming and development, to accommodate essay writing or examining philosophical questions. This way, the AI-Atlas ensures educational objectives for professional and methodical competences on levels K1–K4 [37] by presenting AI and ML as socio-technical and not purely technological. One example of how broader theory can be consolidated by practice are the gamification elements provided in one lab of the AI course (cf. Appendix B). In addition, programming skills are only a means to an end in AI and ML labs, while problem analysis and experimentation become the focus, thereby encouraging exploration.

2.5.2. Fostering Reflection

We suggest to repeatedly ask students to reflect on their preconceptions of the course content and the technical and societal ramifications of this prior knowledge. Making them
think about the myths and ethical problems of the application of AI and learning algorithms
starts an active, though non-visible, process in each individual. For example, in the context
of the ML course, we repeatedly highlight the cognitive dissonance between the current
focus on deep learning methodology in public opinion and the irrefutable results of the
“no free lunch” theorem \[38\]. In the AI course, a lab assignment at the semester start asks
students to create a blog post that presents their well-founded and reflected opinion on the
contents of a futuristic essay \[6\]. At the end of the semester, students can reflect on their
initial statement with a second blog post that may incorporate insights gained throughout
the course. While all opinions are welcome, the emphasis in grading is on self-reflection
and reasoning.

As a more regular intervention, in the AI course, lectures end with an outlook called
“where’s the intelligence?”. It explains why what was discussed is a “clever solution”, but
also what separates it from human-like intelligence. In the ML course, the same time slot is
used to show state-of-the-art implementations of the discussed material. This not only aids
to demystify the technology; it also helps the students spot the kind of tasks they might
approach in their future job using the conveyed foundations.

2.5.3. Encouraging Self-Responsibility and Motivation

Up to twenty percent of the final grade should be acquired by each student during the
semester through self-chosen lab assignments, with results depending on an oral defense
of one’s own work. The choice of 20% is justified by allotting, in this way, a substantial
part of the final result to active participation throughout the semester without replacing
competence-based results with effort-based grading.

From the existing six assignments in the AI course, for example, that are distributed
evenly throughout the semester, students can choose any two to get graded within a
short colloquium between the student team and lecturer during the in-class time (students
usually work on all assignments, but put considerably more time into the two graded ones).
This way, students get empowered to prioritize own learning goals and take ownership
of their investment of time and its distribution over the semester. Even if grading is not
explicitly tied to the lab assignments as in the ML course, the further questions presented
there, the inquiring of the lecturers, and not least the relevance of practical implementation
skills for the final exams motivate students to work deeply on the assignments, even if
sample solutions are freely provided.

A second method to encourage self-exploration and motivation is to set up the labs
in such a fashion that it requires students to independently dive deeper into respective
methods to find practical solutions. The lab descriptions of the ML course, for example,
actively encouraged this, and the lab exercises are often not solvable without going beyond
the lecture content.

2.5.4. Promoting Cooperative Competence Development

Lab assignments should usually be worked on in teams of two students. This way,
students can strategically pair up their existing competencies, as well as learn from each
other. Teams should be allowed to help each other as long as any help is disclosed (accord-
ing to good scientific practice), and competitive elements, such as the public leaderboard
for the AI lab assignment presented in Appendix B, only increase the appeal of and the
necessity for good team work.

2.5.5. Activation of Students

Each 90-min lecture block should contain a part of up to 30 min that assigns an active
role to the students rather than the lecturer. Technical understanding is deepened by
embedding interactive parts, such as small group research tasks and discussion, as well
as thinking and pen-and-paper exercises, thereby increasing the practical treatment of
the subject. For example, in the AI course, a classical brain twister \[39\] can be used to
show the difference between AI (having a computer program appear intelligent) and human
intelligence: approaching it by efficient search through all combinations of possible solution steps constitutes an excellent AI solution for that problem but typically gets labeled “just brute force” by the students at first sight. Other activations in the two exemplary courses take on the forms of jointly solving a puzzle (e.g., “escape from the Wumpus world”), computing results in small groups (e.g., “help inspector Clouseau to probabilistically convict a murderer”), individually applying learned principles (e.g., logic training), or sharing insights from individual research at tables (e.g., exploration of possibilities with OpenAI Gym [40]).

2.5.6. Enabling Social Learning

A prominent place throughout a course based on the AI-Atlas concept should be given to the research work and careers of course alumni and junior teaching staff. Linking course content to concrete outcomes of applied research projects with regional industrial partners known by the students creates a pull that contributes to the students’ motivation and expanded vision for AI and ML in practice, as well as their role in it. Key to create this are tutors (e.g., graduate students) that teach part of the labs: closer in age and role to the course participants, they are in our experience approached frequently by the class to give a second opinion on the more philosophical and career aspects of AI. Innumerable lunches, coffee invitations and after-work beers have been realized this way between teaching staff and students in the AI course and ML course.

2.5.7. Providing Open Educational Resources (OER) and Blended Learning

All course materials, including lecture recordings, slides, and lab materials, should be fully and freely available online [41]. This should enable flexible deepened learning (e.g., for exam preparation), but does not compromise live lecture attendance in our experience. Students can also recap all details when needed on the job as all material is permanently and openly available. This enhances the atlas of AI and ML solution strategies they know by heart. As an add-on, it supports the transition to live online teaching (as was required during the COVID-19-induced lockdowns in 2020, see Section 3.2) as content is already designed to be streaming-friendly.

2.5.8. Creating Practical Career Relevance

Students’ diverse professional backgrounds should be addressed by showing how different AI and ML methods serve as puzzle pieces in numerous everyday situations of (software) engineering. For example, in the AI course and ML course, the lab tasks and in-class exercises are strictly sourced from practical applications, such as automatic university time tabling, biometric access control, or data analysis, to reinforce this point. By connecting the practical coursework with typical tasks of an engineer, programmer or consultant, students clearly see how learning foundations of AI and ML makes them better in their original career goal. By confronting them with new opportunities in and through AI and learning algorithms in business and research, they recognize new and viable career paths (e.g., data scientist [10]) that only begin to gain traction in public awareness.

Additionally, we suggest to invite specialists, ideally course alumni, from regional industrial partners for guest lectures to report on recent successes. Learners should be encouraged to actively use these opportunities to network and engage with those speakers and their ideas. In contrast to the culturally typical reticence of engineering students, the fresh setup with people on stage that might be considered peers age-wise opens them up in the direct (active participation) and metaphorical sense (opening up to the idea of other career options within the fields of AI and ML).

3. Implementation and Analysis

In this section, we evaluate to what extent the different measures advised by the AI-Atlas didactic concept, as implemented within our two exemplary courses, achieve their aspired goals of contributing to providing foundational AI education in times of a height-
ened AI surge. By this, we aim to shed light in a post-hoc fashion on the merits of the AI-Atlas concept per se to serve as a basis for designing future courses for similar needs.

3.1. Participants and Data Basis for Analysis

The principles and settings laid out in the AI-Atlas didactic concept emerged in parallel to designing our exemplary courses out of teaching experience and got codified only upon their completion. The presented analysis, thus, is necessarily a post-hoc analysis: no baseline data from before the AI-Atlas is available due to no existing previous courses on the subject at the involved universities. Furthermore, the analysis is purely based on routinely available data for any course: qualitative student feedback (Likert-scaled, as well as free-text comments) from surveys conducted by the central program administration, and quantitative results from end-of-semester exams (points and graded per task and overall). This is also due to the evolving fashion of our design that had not been planned as a study from the beginning. It means, inter alia, that no control group is available. While this form of evaluation leaves certain aspects to be desired, it is not an uncommon situation for data scientists to having to work with the data that is available in the best possible way, without the possibility to change the basis for evaluations [42]. In what follows, thus, we evaluate our concept with the mindset of data scientists, aiming to establish a relationship between learning success and the employed measures from the AI-Atlas concept.

Demographic and background information on the participants in our courses is listed in Appendix A, while information on specific class sizes (or number of students who returned a questionnaire) will be listed in the captions of the figures below that deal with them. For the ML course, most of these students predominantly seek a career in their original fields of study, though a growing minority considers a job related to ML engineer or ML researcher (the possibility to, e.g., take up graduate studies, is typically completely unknown to our students due to the setup of a “Fachhochschule” [43]). Our students of computer science in the AI course usually envisage a career in software engineering, not specifically AI.

We will be using the qualitative student feedback in Section 3.3 and exam-data in Section 3.4 to evaluate whether the design choices of the AI-Atlas led to a appreciable impact. We do this for two iterations: The first iteration was executed with on-site teaching methods as anticipated by the original AI-Atlas concept (data basis: qualitative and quantitative data from fall term 2019 for the AI course; qualitative data from spring term 2019 for the ML course). The second, due to the COVID-19 pandemic, was executed using hybrid and online teaching, which was not specifically anticipated in the design of the AI-Atlas (data basis: quantitative data from fall term 2020 for the AI course; qualitative data from fall term 2020 and quantitative data from spring and fall terms 2020 for the ML course). The context for the second iteration, the natural experiment that the COVID-19 pandemic provided us with, is described below in Section 3.2. For reasons explained there, we organize our data in the remaining sections as follows: the qualitative feedback is combined for both iterations and evaluated per didactic principle or setting from Section 2. The exam-data is presented specifically per term to allow for comparison between the two iterations.

3.2. Going Online by Necessity

The AI-Atlas was designed specifically for the on-site teaching of small to medium groups (30 to 60 students), but the COVID-19 pandemic forced its execution as hybrid or fully online teaching for two full semesters. Of course, the rapid digitization of higher education in the wake of the COVID-19 pandemic forced teaching around the world to move online in a matter of days. Good teaching methodology differs whether one is going to teach on-site or online [44,45], but the new didactic credo seems to be the one of flexibility: one year into the pandemic, there is still a huge degree of volatility about the possibility, desirability, and potential timeline of returning to a teaching mode of choice, in many countries.
Thus, it is important to know how a course specifically designed for, e.g., on-site teaching, will perform in a hybrid or online setting in terms of the students reaching that course’s educational objectives. This transcends the question of whether going online is merely possible at all [46], as the didactic concept and respective teaching material cannot be adapted that quickly. The move away from on-site teaching was done very rapidly and involuntarily, with the side-effect of no planned, controlled data collection on student learning before and after. Furthermore, as the courses are single-semester, we also do not have longitudinal data within the same cohort of students. Thus, we saw ourselves forced to stray from DBR principles and to (i) combine the qualitative data with (ii) use the quantitative exam data to specifically compare between the effectiveness of the AI-Atlas for on-site and online teaching.

3.3. Qualitative Assessment

In the following subsections, we collect evidence for and against the effectiveness of specific dimensions (i.e., didactic aspects and settings, cf. Sections 2.4 and 2.5) of the AI-Atlas by providing quotations taken from students’ feedback forms at the end of different semesters, and drawing conclusions from them. A short tag at the end of each statement (“AI” or “ML”) indicates the source course. These written qualitative comments are optional for the students to make and, thus, are normally very sparse (though most precious for the improvement of the curriculum and course). The answers might be highly skewed since we cannot control the subset of students that did write some comments. We, therefore, refrain from a statistical analysis of these comments. Nevertheless, the presented example statements are chosen to be representative to allow for the conclusions we draw. In case of counter evidence, we rather over-sample critical comments to avoid any cherry picking (cf. Section 3.3.5).

3.3.1. Dimensions “Canonization and Deconstruction”

The following statements are taken from student’s free and voluntary comments in evaluation surveys:

“Sustainable technologies are taught; in the process you are brought down to earth.” (AI)

“[The] module gives a good overview of the overall topic.” (ML)

“I welcome that […] “where’s the intelligence?” is answered [in] each lecture.” (AI)

Students seem to grasp that AI and ML are ways to solve complex practical problems rather than theories to explain how we think or create artificial life. The content is indeed perceived as a foundation for practice rather than a narrow specialization (cf. Sections 2.4.1 and 2.4.2).

3.3.2. Dimensions “Motivation and Social Learning”

Quality assessment of the two courses is unfortunately neither done in the same way nor regularly by the involved central administrations (cf. Sections 2.5.3 and 2.5.6). In addition to being given the opportunity for providing free-text comments, the students of the ML course are asked to rate their consent with the following statements on a Likert scale (cf. Table 1 for details): (i) Motivation: I regard the lecturer as being motivated and committed; (ii) Competence: I regard the lecturer as being competent in their subject; (iii) Teaching Skills: I regard the lecturer as having good teaching skills; (iv) Clear structure: His/her teaching is clearly structured (a clear thread), and the subject matter was imparted in a comprehensible manner. Table 1 summarizes the evaluation of teaching skills and motivation for the ML course (unfortunately, no evaluation of the ML course was carried out in spring term 2020 due to COVID-19-induced stress in the administration, and the AI course evaluates slightly different questions not pertaining to the dimensions discussed here). The presented score is the averaged score for two lecturers. It mainly reflects the qualitative judgements given by the students also in the free-text comments:
Table 1. Teaching evaluation scores for the ML course. Two teachers were assessed, and the average score is presented (spring 2019: \( n = 30 \), and fall 2020: \( n = 27 \)).

<table>
<thead>
<tr>
<th>Topic</th>
<th>Semester</th>
<th>1 (Strongly Disagree)</th>
<th>2 (Disagree to Some Extent)</th>
<th>3 (Agree to Some Extent)</th>
<th>4 (Strongly Agree)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivation</td>
<td>Fall 2019</td>
<td>1.7%</td>
<td>5.6%</td>
<td>18.8%</td>
<td>79.5%</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>Fall 2020</td>
<td>0.0%</td>
<td>1.9%</td>
<td>13.0%</td>
<td>85.1%</td>
<td>3.7</td>
</tr>
<tr>
<td>Competence</td>
<td>Fall 2019</td>
<td>1.7%</td>
<td>0.0%</td>
<td>15.0%</td>
<td>83.3%</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>Fall 2020</td>
<td>0.0%</td>
<td>1.9%</td>
<td>13.0%</td>
<td>85.1%</td>
<td>3.8</td>
</tr>
<tr>
<td>Teaching skills</td>
<td>Fall 2019</td>
<td>1.7%</td>
<td>15.0%</td>
<td>45.0%</td>
<td>38.3%</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>Fall 2020</td>
<td>3.7%</td>
<td>14.8%</td>
<td>38.9%</td>
<td>42.6%</td>
<td>3.2</td>
</tr>
<tr>
<td>Clear structure</td>
<td>Fall 2019</td>
<td>3.4%</td>
<td>13.4%</td>
<td>48.3%</td>
<td>34.9%</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>Fall 2020</td>
<td>5.6%</td>
<td>14.8%</td>
<td>44.4%</td>
<td>35.2%</td>
<td>3.1</td>
</tr>
</tbody>
</table>

“The professors […] enthusiastically explained it very precisely. I also had the feeling that the fun of the topic seemed very important to them. It was also important for them that everyone understood.” (ML)

“The two lecturers are very motivated and they pass on their enthusiasm and experience in the respective field. I find the exercises and tools we use (Jupyter notebooks, scikit-learn, Orange) very useful and they complement the lessons well. I also appreciate that discussions among each other and in plenary are stimulated.” (ML)

“You can feel that the lecturer is convinced of the subject. He also often brings good examples to help the students on their way.” (AI)

“Very good commitment, super presentation style. Enthusiasm for the subject is obvious and motivates me a lot.” (AI)

Students’ perception of the course contents is in our perspective strongly connected with and dependent on the person that teaches. Insofar, the concept, curriculum or OER availability alone is no guarantee for the intended outcome: enthusiastic teaching is an integral part of the AI-Atlas as it facilitates activation.

3.3.3. Dimension “Activation”

Table 2 summarizes evaluations of the AI course as it has specific questions on the perceived activation (cf. Section 2.5.5) of students and the practical relevance of the presented material (unfortunately, no such questions are asked for the ML course in the central questionnaires as they are issued by a different program administration). Specifically, the students were asked to rate their agreement with the following statements: (i) Activation: the students are actively involved in the teaching process; (ii) Practical relevance: in class, theory is reinforced with examples and applications.

Table 2. Evaluation of the activation in and the practical relevance of the AI course in fall 2019 (unfortunately, no evaluation of the course took place in fall 2020 due to anti-pandemic measures). The number of students who handed in the questionnaire was \( n = 24 \).

<table>
<thead>
<tr>
<th>Topic</th>
<th>Semester</th>
<th>1 (Strongly Disagree)</th>
<th>2 (Disagree to Some Extent)</th>
<th>3 (Agree to Some Extent)</th>
<th>4 (Strongly Agree)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activation</td>
<td>Fall 2019</td>
<td>0.0%</td>
<td>20.8%</td>
<td>29.2%</td>
<td>50.0%</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>Fall 2020</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practical relevance</td>
<td>Fall 2019</td>
<td>0.0%</td>
<td>16.7%</td>
<td>25.0%</td>
<td>58.3%</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>Fall 2020</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These findings are also supported by the following optional written statements that these students handed in:

“The labs support the learning process very much; similarly helpful are the exercises throughout the lectures.” (AI)

“Very handy are the labs where one implements hands-on what should be learned.” (AI)
“Good lecture-style presentation, active presence of the lecturers during the labs that motivates students to listen even on Friday afternoons.” (AI)

A similar response was received from students of the ML course (spring term 2019):

“Good, guided (tutorial) exercises that were very helpful, especially for less advanced students.” (ML)

“The lectures are very interactive.” (ML)

We have increased the time for in-class exercises and interactivity over the years and received increasingly positive feedback on its effects. Despite the success of more modern teaching styles, such as “flipped classroom” [47], lecture-style teaching still seems to be a very helpful didactic setting for technical education if mixed with practical and interactive aspects where applicable.

3.3.4. Dimension “Open Educational Resources”

“The videos on YouTube are ideal for repeating.” (ML)

“The recording of the lectures is very helpful. It gives the students the possibility to review parts of the lecture for exam preparation or if you haven’t understood everything during the lecture.” (ML)

Students use video recordings as intended for repetition without getting distracted by the new flexibility (a real danger of digital transformation: procrastination due to everything being available anytime) (cf. Section 2.5.7).

3.3.5. Criticism: Dimensions “Self-Responsibility and Activation”

Most criticism that we face concerns the workload, the practical work in the lab sessions and the depth of mathematical derivation versus pure application of taught algorithms (cf. Sections 2.5.3 and 2.5.5).

“More exercises during the lecture or in the lab sessions. The topics are not always easy, and small exercises help to learn them correctly.” (AI)

“[The] labs are unfortunately a bit too time-consuming.” (AI)

“The theory necessary for the labs was partly a bit postponed, ... makes the beginning a bit difficult.” (AI)

“Maybe the lab sessions should not be so specific, but should cover a wider range of knowledge and not go into so much detail.” (AI)

“The way the lecturers address the subjects, in my opinion, is very theoretical. There are a lot of mathematical demonstrations that I consider to be out of the scope, and this time should be dedicated to make more examples of the subject. […] Also the course demands way more time than the one available by a full-time student.” (ML)

“Unfortunately, the material covered is just too much. […] you don’t really see the learning objectives for the exam. It’s good to also cover topics superficially and if you need them you can look them up more precisely. But in the lecture, it is not very clear which topics are such.” (ML)

“More exercises, clear knowledge points, mock exam.” (ML)

Table 3 summarizes the evaluation of the appeal and the organization of the ML course for spring term 2019 and fall term 2020 (the same data is not available for the AI course due to differing questionnaires). The students were asked to evaluate the following statements (the same scale applies as seen in Table 1): (i) Labs: the labs supplement the lectures in a meaningful manner and support the learning process; (ii) Material: the support materials (e.g., recommended books, documents handed out) are appropriate; (iii) Organization: the module is sensibly organized, and the coordination between the different lecturers works well.
Table 3. Appeal and organization evaluation scores for the ML course (spring 2019: $n = 28$, and fall 2020: $n = 27$).

<table>
<thead>
<tr>
<th>Topic</th>
<th>Semester</th>
<th>1 (Strongly Disagree)</th>
<th>2 (Disagree to Some Extent)</th>
<th>3 (Agree to Some Extent)</th>
<th>4 (Strongly Agree)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labs</td>
<td>Spring 2019</td>
<td>7.1%</td>
<td>21.4%</td>
<td>50.0%</td>
<td>21.5%</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>Fall 2020</td>
<td>0.0%</td>
<td>14.8%</td>
<td>37.0%</td>
<td>48.2%</td>
<td>3.3</td>
</tr>
<tr>
<td>Material</td>
<td>Spring 2019</td>
<td>0.0%</td>
<td>21.4%</td>
<td>50.0%</td>
<td>28.6%</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>Fall 2020</td>
<td>0.0%</td>
<td>4.2%</td>
<td>41.7%</td>
<td>54.1%</td>
<td>3.5</td>
</tr>
<tr>
<td>Organization</td>
<td>Spring 2019</td>
<td>3.3%</td>
<td>16.7%</td>
<td>60.0%</td>
<td>20.0%</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>Fall 2020</td>
<td>0.0%</td>
<td>25.9%</td>
<td>40.7%</td>
<td>33.4%</td>
<td>3.1</td>
</tr>
</tbody>
</table>

From Table 3, we see that there is much room for improvement. We are of the opinion that learning should be done based on examples and not by pure mathematical derivation. In addition, in a master course of applied science as the ML course, the practical application of the methods should be the deciding factor, and only the necessary mathematical definitions and derivations should be presented. From these critics, we conclude that there should even be more time spent on the application, on lab sessions and exercises. The topics covered in the ML course should be reduced to even less knowledge islands within the big ocean of AI and ML. The lecturer needs the courage to omit certain topics (some islands) and feel confident that the educational objective of generating a helpful map in each student’s mind still can be reached. Over the years, hence, we moved more and more content from the actual lecture slides to the (optional) appendix of the lectures.

3.4. Quantitative Assessment

The following quantitative assessments focus on evaluating student learning success (i.e., reaching of educational objectives) under changing teaching and assessment modes (e-assessments since spring term 2020, on-site written exams before, corresponding to the distance and contact teaching modes during respective terms). In the absence of more direct data to measure the effectiveness of the AI-Atlas, we aim at drawing conclusions about its effectiveness in different teaching and learning settings from these outcome-based comparisons. This is possible since the respective exams where all designed with the same educational objectives and assessment goals, as well as similar question types in mind, irrespective of the rather drastic changes in assessment mode (from close-book in-class to self-supervised, open-book, open-internet).

3.4.1. AI Course Fall Terms 2019 vs. 2020

The content of the AI final exam has stayed largely stable over the years. It contains free text questions used to test students’ ability to precisely define and argue for specific viewpoints; multiple choice exercises to show comprehensive knowledge; programming exercises inspired by the labs; and design tasks with transfer components to reveal higher level competencies. The move from a closed-book written exam to an open-book, open-internet e-assessment in the fall term 2020 of course changed the wording on numerous questions, but the overall depth and composition remained the same, as did the numbers and topics of tasks.

The distributions of the achieved relative scores for pairable tasks as shown in Figure 2 are very similar, which suggests that AI-Atlas suggestions also work well in distance learning mode. Nevertheless, significant differences are discernible in the overall result. The distribution is shifted to the left by about 10 percentage points as shown in Figure 3 (left), as discussed later in Section 4.1. It is striking that, in the planning task (in the bottom left of Figure 2), which is a modeling task involving transfer based on a suitable problem formulation using the Planning Domain Definition Language (PDDL) [48], many students did not know at all how to deal with it in 2020, while, in the group with contact instruction, this rather dry matter could be conveyed and apparently learned well.
The total number of participants was better) for the different tasks of the final exams of the AI course of fall terms 2019 versus 2020.

Figure 2. Histograms and kernel density estimates (kde) of the achieved relative scores (higher is better) for the different tasks of the final exams of the AI course of fall terms 2019 versus 2020. The total number of participants was $n_{2019} = 91$ for fall term 2019 and $n_{2020} = 113$ for the fall term 2020. An overview of the group sample can be found in Table A1. The content of the 9 tasks is irrelevant for the evaluation here unless otherwise noted, but indicated above each plot.

Figure 3. Comparisons of the total achieved relative scores of the AI (left) and ML (right) course exam results of respective semesters (higher is better; kde=kernel density estimate): spring terms 2019 (total number of participants $n = 91$) and 2020 ($n = 13$) for the AI course and spring ($n = 68$) and fall term 2020 ($n = 62$) for the ML course. An overview of the group sample can be found in Table A1 (AI) resp. Table A3 (ML).
3.4.2. ML Course Spring Term 2020 vs. Fall Term 2020

Due to the COVID-19 pandemic, the final assessment of spring and fall terms 2020 had to be taken in full distant mode over the Moodle [49] learning platform. This opened up the opportunity to re-design the existing exam: open book, as online proctoring could not be extended far enough to meaningfully control the use of only permissible aids; and involving hands-on programming, as every participant would sit in front of a well set-up developer’s machine (the personal laptop). For this reason, programming, which was paramount for the lab exercises, could now also be included respectively in the exam in form of two programming tasks, together making up 50% of the exams’ content. This is also the reason why a comparison with pre-pandemic results is not possible for the ML course: the respective exams would be too different to be comparable in a meaningful way.

Participants uploaded Jupyter notebooks [50] containing all programming at the end of the 120-min exam. The programming tasks asked the students to implement a small, but full ML process in Python (using scikit-learn), including (i) explorative data analysis (EDA), (ii) data preprocessing, (iii) feature generation and selection, (iv) algorithm selection, (v) hyper-parameter tuning, (vi) performance assessment, and (vii) comparison and conclusion. Although the tasks were different in topic and based on different data sets, we think that the results, nevertheless, can be compared as they are based on the same learning objectives and lecture parts. With these two programming tasks, we aimed to reach levels K3 and K4 of Bloom’s taxonomy [51] and to test the educational objective of being able to apply and to reflect the ML process on a real data set from end to end.

The result is noteworthy: most of the participants did very well in programming. Figure 4 shows the histograms of the relative scores of the programming tasks. They are left-skewed, meaning that most of the students know how to apply machine learning to solve tasks in real life (the many 0-point entries might be the result of time problems with the exam as a whole, as this was the last task). This indicates that the overall educational objectives of the ML course—to apply ML algorithms—are met by the majority.

Figure 4. Histograms and kernel density estimates (kde) of the achieved relative scores (higher is better) for the programming tasks in the ML course’s spring and fall term final exams 2020. The number of participating students was \( n_{\text{spring}} = 68 \) for the spring, and \( n_{\text{fall}} = 62 \) for the fall term. An overview of the group sample can be found in Table A3.

4. Evaluation and Reflection

According to our experience and as demonstrated above, the AI-Atlas is highly effective for teaching AI and ML principles in an on-site teaching setting. Aside from the presented data here, there are two additional, anecdotal pieces of evidence supporting this claim further: First, our alumni and alumnæ have produced several award-winning theses inspired by the courses and many have ongoing (research) careers in ML. Second,
the AI-Atlas was recognized by the Zurich University of Applied Sciences with the “best teaching—best practice” award in 2019.

4.1. Tracing Weaker Quantitative Results in Online Teaching Mode

In our opinion, one reason for the effectiveness of the AI-Atlas is that it embodies the general learning settings, as shown in Section 2.3, under which students learn best, adapted specifically to the problems faced by current AI and ML tertiary education. In distance teaching mode, this effectiveness suffers somewhat for the AI course as can be seen Figure 3 (left). We performed a discrete Kolmogorov-Smirnov hypothesis test [52] for the total scores of the final exams of the AI and the ML course to check whether the samples of the two semesters stem from a common distribution (Null hypothesis $H_0$). With a significance level of $\alpha = 5\%$, we had to reject the Null hypothesis $H_0$ and assume that there is a significant difference in the distribution of the final scores in both courses when going from either hybrid to online (ML course, just barely different distributions) or from contact to online teaching (AI course, very clearly different distributions). We believe there is one main reason for this drop-off that will need addressing in future iterations, especially in those educational objectives that are based on practical implementation (programming, labs), social interaction (by discussions, competition, study-groups), and the teachers’ presence (theoretical foundations), as explained next.

Tracing the shift in overall grades in the AI course between 2019 and 2020 as visible in Figure 3 (left), we see from Figure 2 that the reduction of effectiveness is mainly due to tasks 1, 5, and 7. These are tasks concerning precise definitions (1) or modeling (5, 7), each with a high transfer aspect. While modeling is technically difficult and complex, precise definitions are a matter of overarching concerns. We hypothesize that these two instances are the first aspects that suffer from the increasingly indirect influence of the lecturer on the students that takes place when shifting from contact to distance teaching: our students are usually technically interested, with less intrinsic motivation for overarching concerns, like, e.g., the precise difference between an AI system and other complex software systems (part of a question on defining what makes up AI). And, while they are technically interested, they are less motivated for prolonged sequences on, e.g., formal logic (as is part of the “planning” topic). In contact mode, the enthusiasm of the lecturer may help carry the motivation of a larger proportion of students through such sequences. In distant mode, where attention is naturally divided between the video stream, chat, the home environment, etc., less motivation is transmitted, and the subjects with the least intrinsic motivation suffer. Additionally, the social learning component through team-work is likely weaker in distant mode with imperfect collaboration tools, so that not every student that would normally be a member of a successful group is able to develop the deep skills necessary for novel modeling tasks on his or her own. Although AI and ML are practically done on a computer, and despite of the fact that “break-out rooms” could be used to organize team work remotely, the social inclusion of each individual likely suffered.

4.2. Tracing Worst Quantitative Results in Hybrid Teaching Mode

Quite counterintuitively, the overall quantitative results for the ML course as depicted in Figure 3 (right) improve again when moving to full online mode. In our opinion, the counterintuitive feeling lifts when considering that it is compared not to on-site teaching in figure 4, but to hybrid teaching. In our experience, hybrid teaching is most demanding for all participants, educators and learners alike, as the teacher has to try to address people in the room, as well as on the computer, which usually results in neglecting one group. In the first pandemic semester of spring term 2020, this condition was worsened by virtually no training for these special circumstances on the side of the educators, and imperfect hardware equipment, leading to frequent technical problems (degraded acoustic quality for discussion in the lecture room, illegible writings on whiteboards for students online, etc.).

We conjecture that the increase in overall scores for the ML course fall term 2020 (full online mode) is due to the ability of the educator in this setting being able to fully con-
centrate on one stakeholder group again, focusing on delivering good streaming content. This interpretation is in line with the somewhat weaker result from the K-S test of the similarity of the two distributions, which are significantly dissimilar, but not as clearly different as for the AI course when going from on-site (good results) to online (degraded results).

4.3. Moving towards a Didactic Concept for Flexible Teaching and Learning

Ultimately, we believe that the AI-Atlas, conceived of and designed for an on-site learning environment, may be equally effective for online teaching if we solve for the problem of transporting some of the more social and teacher-enthusiasm-based design principles in ways that are suitable for the format. This would allow the AI-Atlas to become a fully flexible concept to support teaching regimes that might be any mix of online, hybrid and on-site teaching. What follows are the respective adjustments we are planning for future semesters (regardless of format) that are based on students’ feedback and the principles outlined in Section 2.3 in the light of above’s discussion.

4.3.1. Regarding “Reflection” and “Motivation”

To address the diminished attention span of students and subsequent lower motivation for overarching concerns in distance education settings, we plan more exercises and labs on such aspects to train reflective, interdisciplinary or even holistic thought patterns and reinforce the importance of such “soft” content (cf. Sections 2.5.2 and 2.5.3).

4.3.2. Regarding “Cooperation” and “Social Learning”

To strengthen the connection of students with their core groups for the lab exercises and other key persons in the class (e.g., top students), it is important to solicit quick discussions in the whole video call or frequent, smaller (possibly random) break-out rooms in distance lectures (cf. Sections 2.5.4 and 2.5.6).

4.3.3. Regarding “Activation of Students”

This element can be strengthened by frequent activation of students via online survey tools, like Mentimeter [53], during lectures to have them think on the latest input and produce some output (write a free text answer, make a choice, solve a puzzle or quiz, etc.). We already made positive experiences with a frequency of every second to sixth slide for such pauses for thought (corresponding to a 6–15-min interval between them) (cf. Section 2.5.5).

4.3.4. Regarding “Blended Learning”

The usual best practices for online teaching and learning [54] apply also to AI and ML. What we found especially helpful was to train students to have their cameras on most of the time (increases the perception of connectedness and the degree of interaction), and to roll out tools, like wonder.me (also see Reference [55]), for informal conversation with instantaneous participant groups, much like in a physical break or reception context. This informal networking is important to initiate cooperative competence development, as well (cf. Section 2.5.7).

4.4. Conclusions

With these measures added to the AI-Atlas didactic concept, most of them already implemented within our current AI and ML courses, we see any course crafted after the AI-Atlas principles fit for a very flexible range of teaching modes—whether it is on-site, online, or hybrid. Thus, we recommend the AI-Atlas as a viable basis for consideration when designing tertiary educational courses in AI, ML, and beyond (for the former ones, the syllabi and online teaching material presented in the appendix can also serve as starters to create new respective courses).

Author Contributions: The individual contributions of the authors are as follows: Conceptualization T.S.; Formal analysis C.W.; Investigation C.W.; Methodology T.S. and J.K.; Project administration T.S.; Validation H.G.; Writing (original draft) T.S. and C.W.; Writing (review & editing) J.K., C.W., T.S. and H.G. All authors have read and agreed to the published version of the manuscript.
Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to only standard processes and data sources of the involved study programs being used for data collection and analysis (exams, voluntary feedback).

Informed Consent Statement: Not applicable.

Data Availability Statement: Exam results, voluntary feedback and student inscriptions are hosted at the involved universities in consent with applicable law.

Acknowledgments: We are thankful for the recognition of the AI-Atlas didactic concept through the "best teaching—best practice" award (3rd rank) of the Zurich University of Applied Sciences; for very helpful discussions with and insights on design-based research by Andrea Reichmuth; and for the fruitful discussion on e-didactics with Elisabeth Dumont.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Outline of the AI and ML Modules

Appendix A.1. The AI Course

“Artificial Intelligence 1” (cf. https://stdm.github.io/ai-course/, accessed on 23 June 2021) is a practice-oriented elective course in the final year of a B.Sc. computer science program at a university of applied sciences, encompassing selected foundations of AI and ML and aiming at hands-on problem-solving competency for everyday software challenges. It is geared towards students who have a general curiosity for smartness in software but no aspirations towards research. Most of them, when starting the course, look forward to a career as software engineers, with some thinking about becoming data scientists or about further interdisciplinary studies in areas, like information engineering, speech processing, computer vision, or robotics. This group is quite homogeneous with respect to demography and educational background (cf. Table A1). Age-wise, the students are predominantly in their early twenties, ca. 1–3 years younger as in the ML course due to the AI course taking place at least a year earlier, and some students entering the industry for a while before engaging in master studies. The B.Sc. computer science program can be completed on a full-time or part-time basis.

Table A1. Study model and gender distribution of the students of the AI course from fall term 2019 and fall term 2020.

<table>
<thead>
<tr>
<th>Semester Profile</th>
<th>Fall 2019</th>
<th>Fall 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute</td>
<td>%</td>
</tr>
<tr>
<td>Full time</td>
<td>38</td>
<td>39.6%</td>
</tr>
<tr>
<td>Part time</td>
<td>58</td>
<td>60.4%</td>
</tr>
<tr>
<td>Total</td>
<td>96</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Fall 2019</th>
<th>Fall 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>86</td>
<td>104</td>
</tr>
<tr>
<td>Female</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>

The superior learning objectives are defined as (a) knowing the breadth of AI and particularly ML problem solving strategies, thus identifying such challenges in practice and developing corresponding solutions on one’s own; (b) being able to explain the discussed algorithms and methodologies, thus being enabled to transfer the respective knowledge to the real world. The corresponding syllabus is depicted in Table A2. It is structured in five phases based on the main approaches to AI (symbolic and sub-symbolic) and an elaborate parenthesis dealing with overarching concerns.

The AI course is based on the well-known “AIMA” text book [33] (the much welcomed updates to the recent 4th edition from April 2020 have not yet been adopted; they include a more timely selection and framing of the contents that has partly been anticipated by our curriculum design). It presents AI as a toolbox with separate compartments (=sub-fields), each containing tools to mimic specific aspects of intelligent behavior suitable for certain
ranges of practical problems. The curriculum is special in that it gives equal time to the most relevant ideas from the complete field of AI, not just to fashionable topics around ML and neural networks or the main research areas of the lecturers. The course is taught once per year on-site during fall terms since 2017. The fall term 2020 started in online-only mode and went hybrid for the second half of the course.

Table A2. The curriculum of the AI course, spanning a 14-week semester with 2 lectures and 2 labs (45 min each) per week. On successful completion, the students are awarded 4 ECTS, meaning they have invested ca. 120 h into the coursework (i.e., they spent roughly twice the amount of time in self-study as in class, with most of this time invested into the lab assignments).

<table>
<thead>
<tr>
<th>Topic (Duration)</th>
<th>Key Question</th>
<th>Methods (Excerpt)</th>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction to AI (2 weeks)</td>
<td>What is (artificial) intelligence?</td>
<td>The concept of a rational agent</td>
<td>AI for sci-fi readers: formulating one’s own opinion as a reply to a futuristic essay [6]</td>
</tr>
<tr>
<td>2. Search (3 weeks)</td>
<td>How to find suitable sequences of actions to reach a complex goal?</td>
<td>Uninformed and heuristic search, (Expecti-)Minimax, constraint satisfaction problem solvers</td>
<td>AI for the game “2048”: controlling a number puzzle game (cf. Appendix B)</td>
</tr>
<tr>
<td>3. Knowledge Representation &amp; Planning (3 weeks)</td>
<td>How to represent the world in a way that facilitates reasoning?</td>
<td>Propositional and first order logic, knowledge engineering and reasoning, Datalog for big data, PDDL</td>
<td>AI for a dragnet investigation: finding potential fraudsters using inference over communication meta data</td>
</tr>
<tr>
<td>4. Supervised ML (3 weeks)</td>
<td>What is learning in machines? How to learn from examples?</td>
<td>From linear regression to decision trees and state of the art ensembles</td>
<td>AI for bargain hunters: data mining a dataset of used cars</td>
</tr>
<tr>
<td>5. Selected chapters (2 weeks)</td>
<td>What is the current hype about? How does AI effect society? How could society react?</td>
<td>Primer on deep neural networks and generative adversarial training for image generation</td>
<td>Sci-fi revisited: formulating a reply to the blog post from the first week</td>
</tr>
</tbody>
</table>

Appendix A.2. The ML Course

“Machine Learning” (cf. https://stdm.github.io/ml-course/, accessed on 23 June 2021) is an elective course in an interdisciplinary joint graduate program on engineering of different universities of applied sciences. It builds upon basic knowledge in math, programming, analytics, and statistics as is typically gained in respective undergraduate courses of diverse engineering disciplines and draws on a respective diverse audience with homogeneous demographics (age: 22–25 years) but rather heterogeneous backgrounds (cf. Table A3).

The module teaches the foundations of modern machine learning techniques in a way that focuses on practical applicability to real-world problems. The complete process of building a learning system is considered: formulating the task at hand as a learning problem; extracting useful features from the available data; and choosing and parameterizing a suitable learning algorithm. The syllabus highlights cross-cutting concerns, like ML system design and debugging (how to get intuition into learned models and results), as well as feature engineering, aspects typically cut short in previous courses these students took that touched on learning algorithms.

The corresponding educational objectives are designed as follows: (a) students know the background and taxonomy of machine learning methods; (b) on this basis, they formulate given problems as learning tasks and select a proper learning method; (c) students are able to convert a data set into a trained model by first defining a proper feature set fitting for a task at hand; then they evaluate the chosen approach in a structured way using proper design of experiment; they know how to select models, and “debug” features
and learning algorithms if results do not fit expectations; finally, they are able to leverage on the evaluation framework to tune the parameters of a given system and optimize its performances; (d) students have seen examples of different data sources and problem types and are able to acquire additional expert knowledge from the scientific literature.

Table A3. Background and gender of the students of the ML course from spring term 2019 to fall term 2020.

<table>
<thead>
<tr>
<th>Semester</th>
<th>Profile</th>
<th>Spring 2019</th>
<th>Spring 2020</th>
<th>Fall 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute</td>
<td>%</td>
<td>Absolute</td>
<td>%</td>
</tr>
<tr>
<td>Business</td>
<td>2</td>
<td>3.1%</td>
<td>6</td>
<td>7.6%</td>
</tr>
<tr>
<td>Engineering</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Science</td>
<td>31</td>
<td>39.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer Science</td>
<td>43</td>
<td>67.2%</td>
<td>48</td>
<td>60.8%</td>
</tr>
<tr>
<td>Electrical Engineering</td>
<td>3</td>
<td>4.7%</td>
<td>2</td>
<td>2.5%</td>
</tr>
<tr>
<td>Environmental Science</td>
<td>3</td>
<td>3.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial Technologies</td>
<td>16</td>
<td>25.0%</td>
<td>22</td>
<td>27.8%</td>
</tr>
<tr>
<td>Mechatronics</td>
<td>2</td>
<td>2.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical Engineering</td>
<td>3</td>
<td>3.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aviation</td>
<td>3</td>
<td>3.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geomatics</td>
<td>1</td>
<td>1.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>64</td>
<td>100%</td>
<td>79</td>
<td>100%</td>
</tr>
<tr>
<td>Male</td>
<td>58</td>
<td>90.6%</td>
<td>74</td>
<td>93.7%</td>
</tr>
<tr>
<td>Female</td>
<td>6</td>
<td>9.4%</td>
<td>5</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

The curriculum, depicted in Table A4, spends most time on first principles and illustrates them by specific, selected learning algorithms as the basis for life-long learning in ML. The ML course is not built around any specific textbook, but draws upon multiple sources, including References [33–36,56], having >90% original content. This is contrary to many courses that try to teach a large number of learning algorithms; it also eases the problem of heterogeneous entry competencies where students might have learned about the typical ML algorithms in some class already, but do not know what reasoning led to this specific class of algorithms. The ML course is structured four-fold with an introduction followed by supervised, unsupervised, and reinforcement learning and specifically does not touch neural networks as this is treated in a specialized course. The course has been taught on-site usually once a year in spring terms since 2017. Since spring 2020, the course moved to online teaching mode (with hybrid episodes) and is also taught in the fall term.

Table A4. The 3 ECTS curriculum of the ML course, spanning a 14-week semester of 2 lectures and 1 lab per week.

<table>
<thead>
<tr>
<th>Topic (Duration)</th>
<th>Key Concept</th>
<th>Cross-Cutting Concerns</th>
<th>Methods (Excerpt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction (2 weeks)</td>
<td>Convergence for participants with different backgrounds</td>
<td>Hypothesis space search, inductive bias, computational learning theory, ML as representation-optimization-evaluation</td>
<td>No free lunch theorem, VC dimensions; ML from scratch: implementing linear regression with gradient descent purely from formulae</td>
</tr>
<tr>
<td>2. Supervised learning (7 weeks)</td>
<td>Learning from labeled data</td>
<td>Feature engineering, making the best of limited data, ensemble learning, debugging ML systems, bias-variance trade-off</td>
<td>Cross-validation, learning curve &amp; ceiling analysis, SVMs, bagging, boosting, probabilistic graphical models</td>
</tr>
<tr>
<td>3. Unsupervised learning (3 weeks)</td>
<td>Learning without labels</td>
<td>Probability and Bayesian learning</td>
<td>Dimensionality reduction, anomaly detection, k-means and expectation maximization</td>
</tr>
<tr>
<td>4. Special chapters (2 weeks)</td>
<td>Reinforcement learning</td>
<td>Exploration-exploitation trade-off</td>
<td>AlphaZero</td>
</tr>
</tbody>
</table>
Appendix B. Content Example: AI Model Assignment

This section presents a content example from the AI course in the common format of the community (cf. http://modelai.gettysburg.edu/, accessed on 23 June 2021). It illustrates the technical depth required from the students, as well as the aspects that contribute to the high level of motivation throughout the courses taught according to the AI-Atlas didactic concept.

Appendix B.1. Summary, Topics, and Audience

The lab “2048 game playing agent” (cf. Figure A1) is a four-week assignment at the beginning of the AI course to be approached by pairs of two students (cf. http://stdm.github.io/downloads/courses/AI/P02_2048.zip, accessed on 23 June 2021). It is based on the game “2048” by Gabriele Cirulli (cf. https://play2048.co/, accessed on 23 June 2021) and covers the topics of rational agent development and adversarial search (heuristic search, Expectimax algorithm). The assignment is divided into two distinct phases, each with the task of developing an artificial player that controls the game to win, but different strategies and learning objectives.

Phase one is about taking one’s software development and problem solving skills, together with one’s understanding of the game after a few hours of playing, and implement an agent ad hoc by designing useful heuristics (links to the literature and online forums are provided, where ideas abound). The usual experience of a student after phase one is that it is very difficult and not overly successful to try encoding one’s own strategies purely ad hoc (and that it is impossible to exhaust the knowledge on the web and in the literature without a clear idea of how to conceptually approach the problem).

Phase two introduces the conceptual framework of adversarial heuristic search and the Expectimax algorithm. Students can leverage on their developed ideas of a heuristic function here, but thanks to the look-ahead provided by the search, reach scores usually an order of magnitude higher than their previous results (or manual play). This drives home the point that mapping the problem at hand to the best fitting conceptual/algorithmic approach from the literature pays off way more in AI than investing many hours of manual labor. It also reinforces Sutton’s “bitter lesson” that leveraging compute through search is usually the smartest thing one can do [58].
Appendix B.2. Strengths, Weaknesses, and Difficulty

This assignment’s biggest strength is its addictiveness: students regularly report that they got so caught in the task that they worked through nights and weekends on the hunt for a better high score. This motivation carries over to trying other methods than search: we have seen deep reinforcement learning (RL) approaches developed during these four weeks, despite them not being part of the curriculum. Another strength is its accessibility: students on any skill level find something worthwhile to work on, be it improving their programming skills, understanding a recursive algorithm, or tapping into previously unknown scientific literature to understand RL.

A weakness of the assignment is its dependency on the pace of the corresponding lecture: it helps the educational objective of phase one that the students do not know search algorithms yet (so that they really try ad hoc solutions); it is, however, necessary for phase two that they are acquainted with adversarial search, so that the schedule of the lectures and labs needs to be tightly synced. Another weakness is that much of the initial motivation comes from the students knowing the 2048 game already from its viral history on the web; this effect is fading away over the years.

Appendix B.3. Dependencies and Variants

Platform-independent code templates in Python are given for all technicalities, like interaction with the game, so that students can focus purely on implementing the agent function
\[
\text{next\_move} = f(\text{current\_board})
\]
Students with a good command of any imperative programming language regularly take this as their first attempt to Python programming.

Content-wise, the assignment is preceded by a general introduction to the field of AI, as well as to search algorithms, in the order of one lecture each. Before entering phase two of the assignment, students need to get an introduction to adversarial search and the Expectimax algorithm. An easy variation of the assignment would be to exchange the game by another version that might be more fashionable (and hence able to evoke interest with students) in a few years.

References


17. Siegel, E.V. Why do fools fall into infinite loops: Singing to your computer science class. ACM SIGCSE Bull. 1999, 31, 167–170. [CrossRef]

18. Eaton, E. Teaching integrated AI through interdisciplinary project-driven courses. AI Mag. 2017, 38, 13–21. [CrossRef]


41. Tuomi, I. Open educational resources and the transformation of education. Eur. J. Educ. 2013, 48, 58–78. [CrossRef]

42. Braschler, M.; Stadelmann, T.; Stockinger, K. Applied Data Science; Springer: Cham, Switzerland, 2019.


53. Vallely, K.; Gibson, P. Engaging students on their devices with Mentimeter. *Compass J. Learn. Teach.* 2018, 11. [CrossRef]
55. Gruber, A. Employing innovative technologies to foster foreign language speaking practice. *Acad. Lett.* 2021, 2. [CrossRef]
A Survey of Un-, Weakly-, and Semi-Supervised Learning Methods for Noisy, Missing and Partial Labels in Industrial Vision Applications

Niclas Simmler∗, Pascal Sager†, Philipp Andermatt‡, Ricardo Chavarriaga†, Frank-Peter Schilling†, Matthias Rosenthal†, and Thilo Stadelmann§

∗HSLU Lucerne, Switzerland niclas.simmler@stud.hslu.ch
†ZHAW Datalab Winterthur, Switzerland {sage, chav, scik, rosn, stdm}@zhaw.ch
‡KITRO SA Zurich, Switzerland philipp.andermatt@kitro.ch
§ECLT Venice, Italy Fellow

Abstract—When applying deep learning methods in an industrial vision application, they often fall short of the performance shown in a clean and controlled lab environment due to data quality issues. Few would consider the actual labels as a driving factor, yet inaccurate label data can impair model performance significantly. However, being able to mitigate inaccurate or incomplete labels might also be a cost-saver for real-world projects. Here, we survey state-of-the-art deep learning approaches to resolve such missing labels, noisy labels, and partially labeled data in the prospect of an industrial vision application. We systematically present un-, weakly, and semi-supervised approaches from ‘A’ like anomaly detection to ‘Z’ like zero-shot classification to resolve these challenges by embracing them.

Index Terms—deep learning, computer vision, label quality

I. INTRODUCTION

Utilizing machine learning in an industrial application poses additional challenges compared to research lab environments [1], [2], e.g., in the form of data quality and data quantity issues [3]. “Garbage in, Garbage out” is an often stressed dictum in machine learning – even more so in industrial applications, where data samples and labels collection is difficult and costly [4]. Supervised learning approaches for deep neural networks not only require large amounts of data but also reliably labeled ones. Usually, the data labeling process is conducted manually by individual experts and may involve complex decisions based on years of expert training. To scale-up data labeling in a cost-efficient manner, this process is increasingly outsourced to external contractors (the market for data labeling is expected to reach USD 1.2bn by 2023 [5]). These factors can compromise the labeling quality, often leading to incomplete or uncertain, i.e., noisy labels. Incomplete labels can constitute partially labeled samples or altogether missing labels in a dataset.

To illustrate the importance of clean labels, Fig. 1 shows how model performance is affected by noisy labels. Here, we randomly flipped the labels of a proportion of training data for an exemplary binary classification task in visual quality control. The results show robustness to low proportions of noisy labels (less than 20% noise). However, with increased noise the performance rapidly decreases, reaching a level of guessing if more than 30% of the labels are noisy.

This paper presents a survey of methods aimed at coping with the aforementioned challenges related to data labeling. We focus on methods for computer vision in industrial applications, comprising both image classification and semantic segmentation. The surveyed methods include approaches based on unsupervised, weakly supervised, and semi-supervised learning that lend themselves to real-world applications.

II. PROBLEM DESCRIPTION

The issue of data quality can be caused by factors such as (i) inaccurate or inappropriate sensors causing inconclusive readings; (ii) human mistakes while performing repetitive tasks during data collection; or (iii) financial restriction and too tightly scheduled time-frames for data collections. Frequently, data sample quality (i.e., sensor readings or image quality) is the predominantly optimized part of the data collection process. However, the often neglected aspect of label quality manifests in the three issues of missing, noisy or partial labels. They are the focus of this paper and described in the following paragraphs. Per issue, we present solutions from the following domains, where applicable (see Tab. I):
unsupervised learning (learning from unlabeled data); semi-supervised learning (training on a small subset of labeled data first and subsequently utilizing similar unlabeled data); and weak supervision, which however is an umbrella term [40] for (i) incomplete supervision (e.g., when not all objects in an image are labeled), (ii) inexact supervision (e.g., coarse or loose markings of defects), and (iii) inaccurate supervision (e.g., labels containing mistakes).

**Missing labels:** Resource availability, including financial resources or access to expert labels, constrain real-world projects. This often leads to compromises, e.g., to label only a subset of the data and use machine learning methods to increase the labeled proportion of the data. Various approaches to overcome missing labels are described in Sec. III.

**Noisy labels:** To speed up the collection process, data labeling is often parallelized through the use of multiple experts (or observers). This approach has multiple benefits as (i) it can speed up the labeling process by splitting the work and save time; (ii) depending on the setup, multiple observers can cross-validate results to ensure quality; and (iii) if appropriately applied, multiple opinions on the same sample can mitigate misjudgment and increase trust in the labels. However, this kind of crowdsourcing has its drawbacks since observers may be unreliable and biased [41]. Manual labeling by multiple experts can also cause disagreement in how a sample is labeled (i.e., inter-observer variability [42]), especially when the subject of labeling involves considerable study of the sample. This issue will manifest itself in an uncertainty on the label itself. We cover approaches to resolve respective label noise in Sec. IV.

**Partially labeled data for semantic segmentation:** Fully supervised training of image segmentation models requires pixel-level annotations, i.e., assigning a semantic label (e.g., “carrot”, or “person”) to every single pixel in the image. While detailed pixel-level annotations would yield better models, this is a very time-consuming, hence expensive, process. Alternatively, weak annotations such as scribble annotations [43], point annotations [44], bounding boxes, or image-level labels can be used. Collecting bounding boxes is about 10-15 times faster/cheaper than pixel-level annotation [44], [45]. Image-level labeling, point annotation, and scribble annotation take even less time (around 1-2 seconds per image) [44]. Approaches for reduced labelings are surveyed in Sec. V.

### III. MISSING LABELS

Typical approaches to resolve missing label issues are based on the idea of “finding similar samples”, i.e., contrastive learning (unsupervised), or “label propagation” (semi-supervised).

In the unsupervised contrastive learning [11] approach, a model is trained to discriminate between similar and different images that are all derived from the same unlabeled data by mere augmentation (picking two random images yields a dissimilar pair, taking an image and its augmented version produces a matched pair). In order to classify images using the learned representations, fine-tuning the model using a tiny set of labeled data is required. In recent years, the performance of these methods has increased significantly. The SimCLR [8] framework is able to outperform supervised methods on ImageNet [46]. However, these results may not apply to industrial applications, as they can only be learned with very large batch sizes, i.e., a lot of data and long training times. SimCLRv2 [9] is even larger and more complex. Subsequent knowledge distillation shrinks the model and simplifies deployment. Another approach based on unlabeled data is to utilize a clustering algorithm to group similar features. One of the most recent works in this area is SwAV [10]. This method predicts the cluster assignment of a view from the image representation of another view. Compared to SimCLR, SwAV achieves a slightly higher score on ImageNet. Its disadvantage is the higher complexity, as not only are two views compared, but all of them are clustered.

For a thorough evaluation of semi-supervised learning in the area of image classification, see Ref. [47]. The approaches used are typically split into two categories: (i) the addition of an unsupervised loss term, or (ii) the assignment of pseudo-labels to the unlabeled examples. Popular examples in the first category are the “consistency loss” between the outputs of a network on random perturbations of the same image [12], or the “mean teacher” method [48], which replaces output averaging by averaging of network parameters. The second category uses the regular classifiers to infer pseudo-labels of unlabeled examples by choosing the most confident class [13], [49]. These pseudo-labels are treated like standard labels in the cross-entropy loss. More recently, the “noisy student” training [14] showed improved ImageNet classification performance by training using an EfficientNet [50] model.

Another method to infer the unknown labels is label propagation, where labels of labeled samples are propagated to...
unlabeled samples in close proximity (defined relationally or in terms of similarity). In Ref. [15], it is performed on a large image dataset with convolutional neural network (CNN) [51] descriptors for few-shot learning (FSL). Unseen images are classified via online label propagation, which requires storing the entire dataset while the network is trained in advance and descriptors are fixed. In Ref. [16], label propagation on the training set is performed offline while training the network, such that inference is possible without accessing the original training set. A transductive label propagation method is used, based on the manifold assumption (i.e., that similar examples should get the same prediction), to make predictions on the entire dataset and use them to generate pseudo-labels for the unlabeled data for training. The authors improve the performance on several datasets, especially in the few-labels regime. Ref. [17] proposes a Transductive Propagation Network (TPN) that performs end-to-end labeling of unlabeled images. The network performs the feature extraction using a standard CNN and the graph construction in one. A benchmark on miniImageNet [52] and tieredImageNet [53] shows superior performance compared to other state-of-the-art FSL algorithms especially using zero to five shots, which means it works especially well if only few labels are available. However, a typical issue with FSL is that the training and test samples are disjoint [18]. This causes the feature extractor of a TPN to produce embeddings that are seemingly uncorrelated for unseen classes. This manifests as a disadvantage in terms of robustness when the TPN tries to propagate the labels during graph construction. The Embedding Propagation Network (EPNet) [18] addresses this shortcoming of TPNs by applying the propagation at embedding creation time, thus locating an image’s embedding close to images with similar features in embedding space, resulting in closer labels in their respective space. EPNet achieves superior performance over the TPN architecture in one- and five-shot benchmarking.

The Graph Transduction Game (GTG) [19] is a popular method in the category of label propagation and can be seen as a special case of relaxation labeling [54]–[56], which in turn addresses the problem of label disambiguation. GTG’s general idea is to propagate contextual (i.e., relational) information of labeled instances to classify the unlabeled ones consistently. While, in general, label propagation methods are based on graph Laplacian regularization, GTG is based on non-cooperative game theory. It has been used for the determination of pseudo-labels [57], however in this case, the network is pre-trained, such that the graph remains fixed and there is no weighting mechanism. In general, recent years have seen a steep rise in the application of graph neural networks [58], [59], including graph convolutional neural networks [60] to solve problems which can be represented using graph-structured data. Knowledge graphs can be used as extra information to guide zero-shot classification [61], [62]. The similarity between images in the dataset is also useful in the case of few-shot learning [52]. Ref. [63] proposes to build a weighted fully-connected image network based on similarity and perform message passing in the graph for few-shot recognition. Ref. [64] selects some related entities to build a sub-graph based on object detection results and applies a gated graph neural network to the extracted graph for prediction. Finally, Ref. [65] proposes to build a knowledge graph where the entities are the different categories. The abundance of solutions regarding missing labels promises that this issue can be resolved in real world tasks.

IV. NOISY LABELS

Despite that neural networks exhibit certain robustness towards label noise (cp. Fig. 1 and Refs. [20], [66]), the problem of noisy labels is striking. Various methods in all three learning categories exist to identify and resolve this issue.

Ref. [20] introduces an unsupervised approach while suggesting that noisy labeled samples are harder to learn by a model than correctly labeled ones. This allows for noise identification by fitting a mixture model on the loss values and subsequently using the model’s posterior probabilities to identify noisy labels. Unfortunately, the authors have not yet been able to replicate the modeling approach’s performance on other datasets apart from CIFAR-10 & CIFAR-100 [67]. Collaborative unsupervised domain adaption [21] can mitigate label noise in an unsupervised manner when applied on a real-world dataset. The approach is based on unsupervised domain adaption, which aims to transfer knowledge from a labeled source domain to an unlabeled target domain [22]. The authors evaluate their method by benchmarking it on a medical image diagnosis dataset consisting of H&É stained colon histopathology slides [23] with a convincing performance.

An approach rooted in the area of weak supervision is based on Confident Learning (CL) [24], where the authors do not focus on a particular loss function or model architecture. It uses out-of-sample prediction probabilities that are obtained using cross-validation on noisy labeled data. CL performs (i) the estimation of the joint distribution of noisy labels and true labels, (ii) the identification and pruning of noisy samples, and (iii) the re-training and re-weighting of samples with a new estimated latent prior to identify label noise in a dataset. The main advantage of CL is the absence of hyperparameters and that it does not require guaranteed clean labels. As a result, the authors can test on many publicly available datasets for label noise using CL and, e.g., found an abundant amount of label errors in ImageNet, CIFAR, and even the MNIST dataset [68]. CL is available as an open-source Python package.

Similar to Ref. [20], but in a semi-supervised manner, is the approach followed by the authors of Ref. [27]. The proposed DivideMix architecture models the loss on a sample level using a mixture model to separate clean (labeled) samples from noisy (unlabeled) samples. To prevent confirmation bias, the authors propose to train two networks simultaneously, both generating the sample split for the other network for further training. Both networks then co-refine and co-guess on labeled and unlabeled samples to improve with each iteration. DivideMix achieves outstanding performance on CIFAR-10 and CIFAR-100 and thus outperforms Ref. [20].
V. PARTIALLY LABELED DATA

Here, we survey approaches for learning pixel-level classifications from image-level labels. Some of these approaches formulate the problem as a multi-instance learning (MIL) problem [73], where a multi-class MIL loss is designed for training the network [33], [34]. Generally, weakly supervised approaches try to infer predictions with high information content from labels with low information content. This is especially interesting for semantic segmentation, where generating pixel-level ground truth data is very time-consuming and labor-intensive. Likewise, image-based change detection has similar properties, with the additional challenge of comparing two images and identifying relevant differences.

Ref. [32] uses a directed acyclic graph (DAG) to perform weakly supervised change detection on image series. On top of the DAG, a conditional random field (CRF) model is defined [74] that helps to refine the change mask. Generally, all weakly supervised approaches face the challenge that predictions are coarse. Thus, using a CRF as post-processing is an easy and effective way to increase performance. Ref. [35] proposes a simple architecture that can leverage information from different annotations such as image-level labels, bounding-box labels, and pixel-level labels to improve a semantic segmentation network. The authors use an arbitrary, fully convolutional network to predict the segmentation masks. Afterward, a segmentation mask is fed in an annotation-specific loss module. Depending on the label’s form, a different loss function is applied to improve the segmentation network. They can show that this method can effectively make use of training data with different levels of supervision.

Other approaches use a two-stage approach that generates object-based labels from class activation maps and then trains segmentation networks based on those maps. The class-activation map (CAM) [75] method uses global average pooling, typically applied as a structural regularizer for CNNs, to identify discriminative image regions. This allows using weakly supervised object localization based on networks trained with image-level labels. The coarse response maps are exploited to perform image segmentation using different approaches, including (i) the use of CAMs as the supervisory signal [36], (ii) progressive region refinement based on iterative mining of features [37], or (iii) learning pixel affinity to identify significant regions or propagate pixel-wise information [38], [39]. Further, graph neural networks can be employed to utilize non-local information in the images. In Refs. [76], [77] for example, a Graph-LSTM model is presented to incorporate long-term dependencies together with spatial connections by building graphs and apply the LSTM to propagate neighborhood information globally. Similar ideas have been applied in the case of 3D semantic segmentation and point cloud classification (see e.g., Refs. [78], [79]).

Ref. [31] proposes the weakly supervised change detection method W-CDNet that can be trained with image-level labels. It uses a siamese architecture [80] to compare features from two different images. The change segmentation and classification (CSC) module forces the model to highlight relevant changes. It consists of a custom remapping block that enforces a strong separation between background and foreground pixels, a CRF-RNN layer [81] that refines the change mask, and a classifier that predicts the image-level classification for the image pair. W-CDNet can be trained with both binary image-level labels (describe whether the image contains any relevant change at all) and semantic image-level labels (classify the relevant changes), as well as with full supervision. W-CDNet makes use of existing architectures like U-Net [82] and VGG16 [83], which allows the use of pre-trained weights to speed up the training process. One disadvantage of this method is that the predicted change mask is always a binary mask and not a semantic segmentation mask. Thus, in order to perform semantic change detection, one would have to employ an additional semantic segmentation network on top of the change detector.

This survey suggests that one can achieve similar performance using weaker labels than full (pixel-level) annotations for semantic segmentation tasks.

VI. DISCUSSION AND CONCLUSIONS

Our survey on the issues of missing labels, noisy labels, and partially labeled data in real-world applications of computer vision shows many potential solutions and plenty of active research. However, our survey also shows that there is no silver bullet for either issue: it all depends on the application setting. We conclude by formulating four hypotheses for the further adoption of deep learning in industrial practice:

Shift towards real-world benchmarks: Although various approaches have been tested on standard and artificial datasets, many have not yet seen a noteworthy real-world application. Nevertheless, we could find a few recent candidates in most categories. Overall, this trend [84] in the research community will continue toward more inaccurate and real-world-oriented datasets and
benchmarks is promising and further underlines the importance of the issues covered in this paper. Traditionally, deep learning models rely on high-quality and large datasets. However, some of the presented methods allow for the potential application of these very models on scarce and unreliable data. This paradigm shift allows the introduction of deep learning methods in many more fields, which have been untouched by modern machine learning so far.

**Missing labels sufficiently addressed:** We covered solutions for missing labels in great detail and identified a large amount of promising research work in both the unsupervised and semi-supervised domains. The existence of this large number of potential solutions demonstrates that this specific issue has been considered a significant pain point in the community. Obtaining a large amount of data is hard enough, but labeling a sufficient amount of it is even harder, especially in a corporate setting compared to a global effort of volunteers. The surveyed solutions suggest that this problem can be mitigated.

**Allowing label noise might pay off:** We showed that there are unsupervised, weakly-supervised and semi-supervised methods to counteract noisy labels. Even though noisy labels are rightfully feared when applying deep learning models to real-world data, it is assuring that there are real-world proven methods to overcome the issue. If the intentional admittance of noisy labels in a dataset introduces a quicker turnaround on the data collection process, it can reduce costs and save valuable project time.

**Pixel-level information potentially not necessary:** When dealing with partially labeled data, we presented promising weakly supervised methods. For a real-world application, this implies that it might be easier and faster to label samples using weaker annotations (e.g., scribbles, image-level labels) rather than enforcing exhaustive labeling. Thus, relaxing the labeling requirements can be a considerable cost- and time-saver.

Given the number of potential solutions to crucial but sometimes overlooked (or: underrated) problems in recent years, we expect a greater adoption in deep learning applications in years to come, as the feasibility might incline more and more industries to introduce modern computer vision in their everyday processes.

**ACKNOWLEDGMENTS**

This work is financially supported through Innosuisse grants 36777.1 IP-ICT “FWA” and 26025.1 PFES-ES “QualitAI”.

**REFERENCES**


XGBoost Trained on Synthetic Data to Extract Material Parameters of Organic Semiconductors

Evelyne Knapp
and Mattia Battaglia
Zürich University of Applied Sciences
Institute of Computational Physics
Winterthur, Switzerland
Email: evelyne.knapp@zhaw.ch

Thilo Stadelmann
ZHAW Datalab
Winterthur, Switzerland
ECLT Fellow
Venice, Italy

Sandra Jenatsch and Beat Ruhstaller
Fluxim AG
Winterthur, Switzerland

Abstract—The optimization of organic semiconductor devices relies on the determination of material and device parameters. However, these parameters are often not directly measurable or accessible and may change depending on the neighboring materials in the layered stack. Once the parameters are known, devices can be optimized in order to maximize a certain target, e.g., the brightness of a LED. Here, we combine the use of machine learning and a semiconductor device modeling tool to extract the material parameters from measurements. Therefore, we train our machine learning model with synthetic training data originating from a semiconductor simulator. In a second step, the machine learning model is applied to a measured data set and determines the underlying material parameters. This novel and reliable method for the determination of material parameters paves the way to further device performance optimization.

Index Terms—XGBoost, synthetic data, organic semiconductor, parameter extraction

I. INTRODUCTION

Organic light-emitting diodes (OLEDs) [1] are successfully commercialized in display applications. To overcome limitations in stability and efficiency, further research efforts are crucial. A thorough understanding of the device operation is key which again requires the knowledge of material parameters. Traditionally, material parameters are determined with the aid of dedicated measurements. Some material parameters may vary depending on the sequence of the device layers or measurement techniques, others are not directly measurable. An alternative approach is fitting a device simulation to the corresponding measurement and to derive the material parameters from the simulation. Commonly, least-squares algorithms such as e.g. Levenberg–Marquardt are used to minimize the sum of squared difference between the measurement and simulation. Due to the amount of unknown material parameters and their correlation, multiple (different) experiments [2] are performed and fitted leading to a multi-modal error landscape. The error optimization between measurements and simulations has been demonstrated in various publications [3]–[5], but the process is still not fully automated and requires domain knowledge from the user to direct the search in an appropriate direction or to escape a local minimum. In this contribution, we apply machine learning to the material parameter extraction problem, namely the XGBoost algorithm which is a competitive alternative to neural networks [6]. A set of simplified single-carrier p-doped/intrinsic/p-doped devices varying in thickness is therefore measured and analyzed. The same data has already been used in combination with a manual fitting approach [3]. The production and characterization of such prototype devices is however time-consuming and involves a lot of manual steps. Therefore, the data is extremely scarce and not suited for training a machine learning model. To apply a data-driven machine learning approach a physical model is used for the training data generation.

II. APPROACH AND WORKFLOW

The approach taken in this contribution is visualized in the workflow in Fig. 1. The device under analysis is described in Section III.

Fig. 1. Workflow of synthetic data generation and subsequent machine learning and validation on the measurement data.
As a first step, we generate synthetic data with the aid of a semiconductor simulator. The resulting simulations are current-voltage curves and electrochemical impedance spectroscopy simulations and further described in Section IV. The data set is then split into a training and test set which are firstly used to train the model as explained in Section V and secondly to evaluate the performance on unseen synthetic data as reported in Section VI-A. Once the training is terminated we present the measurements that are of the same structure as the synthetic data to the machine model and show the results in Section VI-B. In this step, the underlying physical material parameters are predicted which are then used in a semiconductor simulation to reproduce the measured experiment. Once a good agreement between the measurement and the simulation is obtained, the optimization process of the device would start and the influence of parameters on the overall performance would be investigated which is, however, not further pursued in this work. The approach introduced above assumes that the physical model captures the main features in the measurements and is an adequate description of the underlying physical processes.

III. DEVICE UNDER INVESTIGATION

The analysis is concerned with three hole-only devices consisting of a 100, 150, and 200 nm thick intrinsic tris(3-phenyl-1H-benzoimidazol-1-yl-2(3H)-ylidene)-1,2-phenylene)Ir (DPBIC) layer, respectively, which is sandwiched between two 30 nm thick, 10 Vol. % MoO3 doped (p-type) DPBIC layers. The contact is made of indium tin oxide (ITO) and gold (Au) which ensure a good band alignment with the Highest Occupied Molecular Orbital (HOMO) of DPBIC (5.28 eV [3]). The device structure is shown in Fig. 2 (not to scale) with an additional external series resistance. All devices were measured at room temperature and in the dark with Paiao [7]. The current–voltage measurement is the most basic characterization method for OLEDs and solar cells. A more advanced technique is the electrochemical impedance spectroscopy that determines the impedance, i.e. the AC resistance, of electrochemical systems as a function of voltage resulting in an impedance-frequency representation. The second impedance measurement uses a fixed frequency while we sweep the voltage while the frequency is swept. The second impedance representation for the impedance at a certain offset frequency is calculated according to $Z = \frac{V_{AC}}{I_{AC}}$. As the current might be phase-shifted with respect to the AC voltage modulation the impedance $Z$ is complex and can be represented in different ways. We will use the Nyquist representation for the impedance at a certain offset voltage while the frequency is swept. The second impedance measurement uses a fixed frequency while we sweep the voltage resulting in an impedance-frequency representation. All measurements on one device were carried out subsequently without changing the contact pins. For impedance analysis, an oscillating voltage modulation of 70mV was used.

IV. SYNTHETIC DATA GENERATION

Algorithms, frameworks and machine learning packages have flourished over the last decade and a wide range and variety is available. A more delicate and scarce resource is high-quality data [8]. To circumvent this problem we generate synthetic data. Synthetic data is artificially created by simulations and not collected from the real world or generated by actual events. The advantages of synthetic data generation are manifold: The amount of data can easily be increased at the cost of simulation time. The diversity of data can be chosen such that all possible scenarios are included and the data is perfectly annotated.

For our data generation case, we refer to semiconductor modelling and use a simplified and reduced OLED structure that facilitates material characterization for holes in organic semiconductors. Further, the three-dimensional OLED geometry is reduced to a one-dimensional simulation domain as shown in Fig. 2 as the red line. The material parameters that determine the behavior of the device are mostly unknown or can only be measured with great effort. The goal of this work is to determine these underlying material parameters from device characterization measurements. We display the unknown parameters in Fig. 2 in their corresponding domain.

![Fig. 2. Device structure and material parameters of a single-carrier device. The red line indicates the one-dimensional simulation domain for the semiconductor modelling.](image)
layers differ in terms of the relative permittivity $\epsilon$ and the doping density $D$ from the intrinsic layer. The field-dependent Poole–Frenkel mobility model, $\mu(E) = \mu_0 \exp(\gamma \sqrt{|E|})$ where $\mu_0$ is the zero-field mobility, $\gamma$ the field-enhancement factor, and $E$ the electric field is the same in all semiconductor layers. As input parameters for the simulation we vary the values within the boundaries indicated in Table 1. Depending on the parameter the values are uniformly or log-uniformly sampled. For the physical interpretation of the material parameters refer to [3]. With the simulation tool Setfos [9] we create, by randomly varying the material parameters, 100’000 sample simulations for all three thicknesses (100, 150, and 200 nm) that are used for the training. Therefore, we solve the system of coupled partial differential equations for semiconductors [10]–[12] on the one-dimensional domain and vary the seven material parameters simultaneously within prescribed boundaries from Table I. In Fig. 3 a single sample of the synthetic data set is shown with the predictor and target variables as well as their sizes. The input data of the physical model is the output data of the machine model and vice versa. In summary, the data set consists of a current-voltage simulation, impedance simulations at two bias voltages with a frequency sweep, and an impedance simulation for a fixed frequency with a voltage sweep for each device thickness.

V. TRAINING OF MACHINE LEARNING MODEL

With the synthetic data we proceed in the flow chart in Fig. 1 and split the data in training and test sets in a ratio of 80% to 20%. The training set is used to train the machine model while we validate the model with the test set. We deal with a multi-target regression problem that is concerned with the prediction of multiple continuous target variables using a shared set of predictors. In the following we apply XGBoost [13] which is short for eXtreme Gradient Boosting package to the regression problem. It was created by Tianqi Chen [14] and is an efficient and scalable implementation of gradient boosting framework by [15], [16] along with some regularization factors. XGBoost is considered for supervised machine learning tasks in classification and regression and has performed well for structured, tabular data in the past.

XGBoost is categorized as boosting techniques in ensemble learning. Boosting is a method that combines simpler and weaker models (trees) to make better predictions of the target variable. Models are added gradually and in a sequential manner until there are no more improvements in the predictions. Gradient boosting uses the gradient descent algorithm to add the simpler models and thereby minimizes a regularized objective function. It is a combination of a convex loss function that takes the difference between the predicted and target outputs into account and a penalty term for model complexity. The training is performed iteratively by adding a new tree to reduce the residuals of the current ensemble of trees.

We create for each target variable an XGBoost model that is individually trained. The target variables were previously transformed by logarithmizing (if necessary) and scaling each target between 0 and 1. The XGBoost model was trained on 80’000 training samples and validated on 20’000 test samples. For the training we use a manually tuned learning rate $\eta = 0.25$, the maximum depth of a tree is limited to 15 and the maximum number of trees to 40. As a loss function the Root Mean Square Error (RMSE) is selected. In Fig. 4 the train and validation loss is plotted against the number of trees for three exemplary parameters. The first material parameter is the zero-field mobility $\mu_0$ and represents a successful training. The training loss function decreases fast and levels out. The loss function of the test set shows a very similar behavior with the difference that the plateau is slightly higher. The second parameter is the intrinsic relative permittivity $\epsilon_i$. Its performance is to some extent worse since the training set reaches a clearly lower value than the test set. The third parameter, the doped relative permittivity $\epsilon_d$, behaves differently because the training loss is still decreasing while the loss function of the test set remains on the same level. The reason for the worst performance lies in the sensitivity of the physical model to the particular parameter, i.e. the model is not very sensitive to the doped relative permittivity $\epsilon_d$ and therefore hard to predict, or in other words, a change in the

![Fig. 3. Simulated predictor (top) and corresponding target (bottom) variables with their dimensions. The physical model works in the opposite direction as the machine learning model.](image-url)
doped relative permittivity $\epsilon_d$ has little impact on the result of the semiconductor device simulation in contrast to other material parameters.

The remaining parameters act as in the first or second case in Fig. 4. Only the work function $\phi$ has the characteristic of the third scenario, the physical reasoning being that the work function $\phi$ has little impact on the device properties if the boundary layers are highly doped.

VI. Evaluation

A. Synthetic Test Data

The performance of the XGBoost model on the material parameters is analyzed again for the three exemplary parameters in Fig. 5 where an identity chart is displayed. The material parameter value predicted by the XGBoost model versus the true parameter value from the training or test set is plotted. As a guide-to-the-eye the red diagonal represents perfect prediction on the identity chart. The first parameter is the zero-field mobility $\mu_0$ which can be very well predicted even for the test set as displayed in a). Also decreasing the size of the training set has only little effect on the prediction. As a second parameter we present in b) the intrinsic relative permittivity $\epsilon_i$ where the prediction is still valid, but worse than in the previous case. The uniform sampling of the relative permittivity was rounded to one decimal place as visible in the figure. A parameter that is more difficult to predict is the doped relative permittivity $\epsilon_d$ shown in c). The spread around the diagonal is wider for the training set and even worse for the test set. The results in Fig. 5 are very much in line with Fig. 4 and the identity charts visualize again the outcome of the training phase.

B. Measurement Data

In this section, we feed the XGBoost model with the measured current-voltage and impedance spectroscopy data. Since the measurements are reliable and repeatable, we have not performed any pre-processing on the data such as e.g. de-noising. The measured data set has exactly the same structure as in the synthetic test data.
TABLE II
EXTRACTED MATERIAL PARAMETERS FROM XGBOOST

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_s )</td>
<td>49</td>
<td>( \Omega )</td>
</tr>
<tr>
<td>( \phi )</td>
<td>5.00</td>
<td>eV</td>
</tr>
<tr>
<td>( \mu_0 )</td>
<td>( 1.1 \times 10^{-7} )</td>
<td>( \text{cm}^2/(\text{Vs}) )</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>( 8.5 \times 10^{-4} )</td>
<td>( \text{V/m}/\sqrt[4]{\text{V}} )</td>
</tr>
<tr>
<td>( D )</td>
<td>( 2.7 \times 10^{-6} )</td>
<td>( \text{m}^{-3} )</td>
</tr>
<tr>
<td>( e_d )</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>( e_f )</td>
<td>4.2</td>
<td></td>
</tr>
</tbody>
</table>

as the synthetic data in Fig. 3 and is fed to the XGBoost model. The extracted material parameters are re-transformed and shown in Table II. We note that these material parameters obtained with the help of machine learning compare favorably with the ones obtained in a more traditional semi-automatic least square fitting approach [3].

Depending on the size of the training set and the selected hyperparameters the extracted material parameters will slightly vary. Increasing the training data set helps to reduce this variation e.g. for the doping parameter \( D \). A hyperparameter search for each material parameter has the potential to further improve the training and find the optimum configuration of XGBoost. In order to circumvent this problem the resulting material parameter set can be further processed and serve as an initial guess to a local optimization algorithm that minimizes the error between the measurements and the simulations.

As a next step, the material parameters are fed back into the semiconductor simulator to reproduce the measurements. In Fig. 6 the measurements as well as the simulations based on the material parameters predicted by XGBoost are displayed. The first figure shows the measured and simulated current-voltage curves with a very good agreement. Also in the second plot the impedance in log-log representation for two offset voltages represents an accurate description. The plots at the bottom display the real and imaginary part, respectively, of the impedance versus the applied voltage at a constant frequency of 28.8 kHz. The characteristic features are captured in all four situations. The final agreement between measurements and simulations confirm that the semiconductor model consists of all important physical ingredients to describe the measurement with one set of material parameters.

VII. CONCLUSIONS

A material parameter extraction problem for single-carrier organic semiconductor devices with three different thicknesses was presented. The approach taken in this work combined a physical semiconductor model for synthetic data generation and machine learning. We successfully trained an XGBoost model on the synthetic data to a multi-target regression problem to determine underlying material parameters from current-voltage and impedance spectroscopy measurement data. The material parameters extracted from the measurement were fed to the semiconductor simulator. The simulation and the measurement are in close agreement. The concept of merging machine learning and physical modelling for data generation is a powerful alternative to classical fitting algorithms provided the simulation times for the physical modelling are short. As a next step, we will introduce a quantitative measure of the fit quality. Further, we envisage to extend the application to the determination of the underlying physical model with its key ingredients. Such a physics-informed machine learning approach can potentially be helpful for various applied physics and engineering problems.

ACKNOWLEDGMENT

Financial support from Innosuisse (project AIPV 37304.1-IP) in the impulse programme digitalisation is acknowledged. P.-A. Will, S. Lenk and S. Reineke from TU Dresden are acknowledged for providing the organic semiconductor devices based on which the measured data was acquired.

REFERENCES

(a) The current-voltage curves for all three thicknesses are shown.

(b) The absolute imaginary part of the impedance is plotted versus the real part of the impedance.

(c) The real part of the impedance versus the applied voltage at a frequency of 28.8 kHz is displayed.

(d) The absolute imaginary part of the impedance versus the applied voltage at a frequency of 28.8 kHz is shown.

Fig. 6. Predicted simulations based on extracted material parameters by XGBoost in comparison with the measured data.
Context–aware learning for Finite Mixture Models

Serafeim Perdikis¹,* Robert Leeb¹, Ricardo Chavarriaga¹, José del R. Millán¹

¹Chair in Brain-Machine Interface, Center for Neuroprosthetics, Institute of Bioengineering, School of Engineering, École Polytechnique Fédérale de Lausanne, Campus Biotech H4, Chemin des Mines 9, CH-1202, Geneva, Switzerland

Abstract

This work investigates a class of algorithms exploiting contextual information to improve unsupervised maximum–likelihood estimation (MLE) of finite mixture models (FMM). These algorithms are derived in a probabilistic setting where the regular FMM graphs can be extended with context–related variables, applying an expectation–maximization (EM) approach which renders explicit supervision completely redundant. We show that, by direct application of the missing information principle (MIP), the algorithms’ performances range between those of the regular supervised and unsupervised MLEs, proportionally to the information content of the contextual assistance. Our simulation results demonstrate the superiority of context–aware FMM learning as compared to conventional unsupervised training in terms of estimation precision, standard errors, convergence rates and classification accuracy or regression fitness in various scenarios, while also highlighting important properties and differences among the outlined situations. The applicability of this approach is showcased in three real–world scenarios.

Keywords: context–awareness, side–information, unsupervised learning, probabilistic labels, finite mixture models, expectation–maximization, maximum–likelihood, parameter estimation

*Corresponding author

Email addresses: serafeim.perdikis@epfl.ch (Serafeim Perdikis), robert.leeb@epfl.ch (Robert Leeb), ricardo.chavarriaga@epfl.ch (Ricardo Chavarriaga), jose.millan@epfl.ch (José del R. Millán)
1. Introduction

The commonly encountered situation of missing data labels has raised an increasing interest in unsupervised learning approaches for classification. Unsupervised classification can be defined as the task of estimating the parameters of a classification model when the number and type of classes are known, training data samples are available, but there exist no associated ground truth data labels whatsoever. The latter characteristic distinguishes this problem from semi-supervised learning methods [1], where some labeled instances exist. The absence of any kind of reward signal renders reinforcement [2] equally unsuitable. Recent works have showcased that, even in this setting, there exist ways to improve parameter estimation over unsupervised techniques by exploiting additional, side-information.

Along these lines, this work studies algorithms which can exploit probabilistic contextual information to improve expectation–maximization (EM)–based, maximum–likelihood (ML) estimation (MLE) in generative finite mixture models (FMM) [3, chap. 9]. More specifically, we focus on situations where it is possible to extend the probabilistic directed graph of FMMs with contextual random variables $c_i$ whose prior, $p(c_i)$, and/or conditional distributions, $p(z_i|c_i)$ or $p(c_i|z_i)$\(^1\), are known, thus providing the additional side–information. We show that such contextual assistance is able to partially reveal the missing data label information.

As illustrative examples, one can consider an adaptive activity recognition system equipped with online unsupervised learning capabilities to classify a set of activity classes $z$ from kinematic sensor data features $x$. Such a system could benefit from environmental context $c \in \{h(ome), o(utdoor)\}$, upon which $z$ naturally depends. That is, since the aforementioned distributions (e.g., $p(z = \text{run}|c = o)$ and $p(z = \text{walk}|c = o)$) are indicative of the current activity even in the case of latent context $c$, which is a consequence of the statistical relationships

\(^1\)Variable $z_i$ represents the latent class label of data sample $x_i$. 
between context and labels, i.e., the facts that running is more likely to occur outdoors, while walking indoors. Similarly, unsupervised learning of lung tumor detectors with \( z \in \{ \text{malignant, benign} \} \) from X-ray imaging features \( \mathbf{x} \), could be enhanced by knowledge on the results of a parallel blood test \( c \) (observed, but conditionally independent from \( \mathbf{x} \)), where the dependency relationship between \( z \) and \( c \) is reversed with respect to the previous example.

The main motivation of this study is to show that such algorithms are able to learn “better” than their unsupervised equivalents and close to the supervised ones despite completely discarding any need for ground truth. Secondly, we wish to explore the information-theoretic principles under which this type of side-information yields estimation benefits. The contributions of this article are threefold. First, we draw attention to the fact that simple EM–MLE along with the above mild assumptions result in improved unsupervised learning, a fact so far neglected in favor of more complex methodologies [4, 5]. Second, we prove this framework’s benefits in various FMM scenarios in terms of parameter estimation precision, standard errors, convergence rates and classification or regression quality. A comparative analysis of the algorithms in question is also offered. Additionally, we demonstrate the applicability of this approach to real–world problems. The third contribution entails the in-depth study of the underlying mechanisms through which these algorithms improve unsupervised estimation. This includes, on the one hand, the analysis of exemplary likelihood landscapes. On the other hand, we explicitly demonstrate—for the first time—the alleviation of missing label information by side–information, through the missing information principle (MIP) [6].

The remainder of this manuscript is organized as follows: Section 2 discusses the relevant literature and highlights its differences with the present work. Section 3 presents the examined algorithms, the relevant theory and the evaluation methodology. Section 4 illustrates the results on artificial and real datasets. Finally, Section 5 discusses the proposed approach in the light of the results.
2. Related work

A great deal of literature addresses various cases of weak supervision where, although some form of data labels is available, it differs from regular supervision. A first case concerns learning from partially or ambiguously labeled datasets, where each data sample is associated to many possible labels only one of which is correct [7, 8]. Second, multi-label, multi-annotator (crowd-sourcing) settings where all of the labels could be valid, potentially with different or time-varying reliability [9, 10, 11]. Partial-label problems, where labels are only missing for some of the classes, are studied in [12]. In [13], another partial-label framework is investigated, concerning the case where one knows to which classes a sample does not belong. Additionally, multiple-instance or multi-view learning methods, where each learning example contains a bag of samples are proposed in [14, 15, 16, 17]. In [18], a generic method to handle most of the above problems is presented. Nguyen et al. [19] put forward a framework exploiting additional information in the form of reliability indices of data labels. Similarly, cases with noisy or wrong labels are addressed in [20, 21, 22, 23]. The setting discussed here differs substantially from all these approaches, as well as from co-training [24, 25] and all other semi-supervised learning methodologies, in that the contextual random variables can be virtually anything, including, but not restricted to some kind of explicit labeling. Hence, side-information on data labels emerges naturally through the dependence relationship between the latent label/class and the contextual variables taking the form of implicit, but not actual “probabilistic labels”. Essentially, our framework proposes how “soft” labels can be derived by context without manual effort and explicit labelers.

Another class of related problems regards those where side-information is provided in the form of constraints. Most of the early work has focused on known positive and/or negative linkage between pairs or sets of samples [26, 27]. Beyond case-specific methods, there exist frameworks able to cope with context-aware learning irrespectively of the form of side-information.

Chang et al. [28] have proposed constraint-driven learning (CODL), which
penalizes constraint violations of a given model by augmenting the objective function with a penalty term. Nevertheless, its formulation assumes labeled instances for initialization, does not maintain uncertainty during learning, and involves a fairly heuristic optimization algorithm with many hyperparameters. Liang et al. [29] put forward a Bayesian approach by modeling side-information as so-called “measurements”: noisy expectations of constraint features. The employed objective function is optimized with a complicated variational approximation which is the method’s main disadvantage.

In a series of articles, McCallum and colleagues have introduced Generalized Expectation Criteria (GEC), where the additional information comes as linear constraints of a set of feature expectations forming a standalone objective or augmenting the common likelihood objective with an extra term [4]. A special case of GEC had been initially proposed as “expectation regularization” [30]. Several optimization procedures have been presented and tested, including gradient descent [31] and variational approximation [32].

Using the very same modeling of side-information, Ganchev et al. [5] have proposed Posterior Regularization (PR). In this case, constraints are imposed directly on the posterior distributions of latent models, giving rise to optimization algorithms akin to regular EM. PR’s conceptual intuitiveness has contributed to its recent popularity [33, 34, 35]. Ghosh et al. [36] have independently proposed a PR formulation specific to FMMs and constraints in the form of a-priori knowledge of mixing proportions, deriving a variant of the “scaled”–PR algorithm for this particular problem [5, Appendix A]. Despite sharing the same model, this work exploits a less generic type of side-information and involves complex formulations.

In a brilliant analysis [5, Section 4], it is shown that under certain approximations all four generic frameworks are equivalent. Compared to these approaches, it can be said that the algorithms examined here trade-off generality in favor of simplicity and intuitiveness. This claim is substantiated in Appendix A, where the PR-equivalents of our algorithms are discussed.

The idea of augmenting a given model to include context, the cornerstone
of our work, can be traced back to the “hierarchical shrinkage” method [37]. Probabilistic context modeling identical to ours is proposed in [38]. However, in this case the authors focus on classification improvements rather than the estimation properties of the algorithm.

Some of the aforementioned studies propose algorithms with identical formulations to those proposed in the present work. Specifically, Bouveyron et al. [20] and Côme et al. [39] have produced the formulation of what we call here the WCA algorithm, in the context of learning with noisy labels and through Dempster–Shafer basic belief assignments, respectively. On the other hand, Ambroise et al. [13] and Szczurek et al. [21] (who also compare to the work of Côme et al.) arrive at the formulation of our CA algorithm assuming, again, the existence of “soft” supervision. Our work is, first, more general than those, since we exhaustively compare all algorithmic possibilities. Most importantly, as already mentioned, our derivations do not take the existence of “uncertain” labels for granted and discard the need for any kind of ground truth. Finally, the scope of this article is the only one strongly focused on the information-theoretic effects of learning with side–information.

3. Methods

3.1. Context–aware learning algorithms for FMMs

In order to gain a solid understanding of the proposed idea, the reader should recall [3, chap. 9.2] that a FMM is represented by the Bayesian network illustrated in Fig. 1a–b (enclosed in a dashed box), where \( x_i \in X \) is the observed independent and identically distributed (iid) data samples of a dataset \( X \) with cardinality \( N \) \((i \in [1, N])\), \( z_i \in Z \) is the latent data representing the mixture/class generating sample \( x_i \) having a 1–of–M representation, so that \( z_{ij} \in \{0, 1\} \), \( \sum_j z_{ij} = 1 \) and \( M \) the number of mixtures/classes. The distribution of observed data \( x \) is then \( p(x) = \sum_z p(x, z) = \sum_z p(z)p(x|z) = \sum_{j=1}^{M} \pi_j f_j(x, \theta_j) \), where, \( \pi_j = p(z_j = 1) \) are the mixture coefficients with \( \sum_{j=1}^{M} \pi_j = 1 \) and \( f(x, \theta') = p(x|z, \theta') \) with \( f \) belonging to some identifiable family with parame-
Figure 1: Graphical representations of augmented (solid boxes) and regular (dashed boxes) mixture models for a set of N independent and identically distributed (iid) data samples. Random variables depicted in circles, transparent for latent variables, shaded for observed variables and stripping for variables that can be observed or latent on occasion. Model parameters are illustrated with squares. \( x_i \) are the observed data samples, \( z_i \) the latent class labels and \( c_i \) the contextual variables. Model (a) gives rise to CA-type of estimation and (b) to WCA.

ML estimation consists in maximizing the logarithm of the incomplete-data, marginal likelihood \( \log L(\theta|X) = \log(\prod_{i=1}^{N} p(x_i)) \) over \( \theta \). In supervised estimation, \( z_i \) are the observed labels \( y_i \), yielding analytic solutions. Conversely, for latent \( z_i \) one relies on the iterative EM–MLE, where, first, the expectation (under posteriors \( p(z|x, \theta) \)) of the complete-data log–likelihood \( \log L_c(\theta|X, Z) \) is formed (E–step):

\[
Q(\theta, \hat{\theta}^k) = \mathbb{E}_{\hat{\theta}^k}\{\log L_c(\theta|X, Z)\} = \\
\sum_{i,j}^{N,M} \mathbb{E}_{\hat{\theta}^k}\{z_{ij}\}\log \pi_j + \sum_{i,j}^{N,M} \mathbb{E}_{\hat{\theta}^k}\{z_{ij}\}\log(f_j(x_i, \theta_j))
\]

where \( \theta = \{\pi_j, \theta_j\}, \forall j \) are the overall estimated parameters and \( \hat{\theta}^k \) the \( k^{th} \) estimate. Then, \( Q(\theta, \hat{\theta}^k) \) can be analytically maximized (M–step): \( \hat{\theta}^{k+1} = \arg \max_{\theta} \{Q(\theta, \hat{\theta}^k)\} \).

This conventional unsupervised EM–MLE algorithm (termed hereafter US) is known to get stuck in local maxima (thus being sensitive to the initialization \( \hat{\theta}^0 \)) and exhibits compromised estimation precision compared to supervised estimation (termed S). Furthermore, it is inferior to S in terms of standard errors and convergence rate (since it is iterative). It is clear that these limitations should be related to the missing label information. Both methods share the
same objective of (1), only differing in the replacement of labels \( y_i \) (\( S \)) by posteriors \( \mathbb{E}_{\hat{\theta}^k}\{z_{ij}\} = p(z_i = j|x_i, \hat{\theta}^k) \) (\( US \)). Hence, it is reasonable to assume that boosting the information content (entropy) of distributions \( \mathbb{E}_{\hat{\theta}^k}\{z_{ij}\} \) towards the labels \( y_i \) should raise \( US \) performances closer to those of \( S \).

The idea put forward in this article is to achieve this goal by directly embedding probabilistic side-information into generative Bayesian networks (directed graphs). More specifically, it suffices that a) contextual information can be modeled by (in general, latent) random variables \( c_i \), which b) can be assumed to have a dependence relationship with the latent nodes \( z_i \) (augmenting the underlying model, as shown in Fig. 1 for the case of FMMs) and c) whose distributions \( p(c_i) \) and/or \( p(z_i|c_i) \), \( p(c_i|z_i) \) are known. Given these prerequisites, deriving context-aware algorithms results from straightforward application of EM on the augmented models. Analytical derivations can be found in Appendix B.

It is critical to discuss what these assumptions imply for the applicability of the proposed approach. The first prerequisite is a mere modeling choice and hardly restrictive, since all natural quantities can be modeled as random variables. The second assumption forms the basis of our framework. It advocates for a paradigm shift where one needs not solely rely on the possibility to collect usual “data and labels”, but can instead identify contextual sources of information that may partially reveal the missing data labels. Of course, this might not always be possible. The third assumption can also be limiting since, even after identifying potentially useful types of context, the distributions \( p(c), p(z|c), p(c|z) \) could still be unknown, difficult to pre-estimate, or rather uninformative. For instance, in the medical informatics example used in Section 1, medical tests additional to X-ray imaging for lung tumour detection might be as hard, expensive or dangerous to collect as the biopsy that would reveal the actual ground truth labels (malignant or benign). In addition to this, this last prerequisite can only be satisfied for fully defined \( z \) (i.e., known number of mix-

\(^2\)Without loss of generality, the contextual random variables will be assumed hereafter to be univariate and discrete.
tures $M$), which limits the scope to unsupervised classification. Hence, with respect to the last two assumptions, there can be no guarantees of existence or benefits, or even a way to analytically quantify the likelihood of those in general applications. Nevertheless, in the era of information explosion and the emergence of the Internet Of Things, we believe that these assumptions can be already satisfied more often than not, with the situation improving in the foreseeable future.

The two possible types of dependence between $c_i$ and $z_i$ give rise to two different augmented models (Fig. 1a and b) and, therefore, two corresponding EM–MLE algorithms termed CA and WCA, respectively. We are also considering a third, heuristic algorithm termed DCA (Direct Context–Aware), where the posterior distribution of latent labels is defined by arbitrary probabilistic labels and the evidence $X$ is ignored. Table 1 summarizes the naming convention, probabilistic labels $p_i$, E-step and log$L$ formulation of each algorithm.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>E-step</th>
<th>$\log L = \sum_{i=1}^{N} \log(\sum_{j=1}^{M}(\ldots))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>None</td>
<td>$\pi_j f_j(x_i</td>
</tr>
<tr>
<td>CA</td>
<td>$p_i, \pi_j f_j(x_i</td>
<td>\hat{\theta})$</td>
</tr>
<tr>
<td>WCA</td>
<td>$p_i, \pi_j f_j(x_i</td>
<td>\hat{\theta})$</td>
</tr>
<tr>
<td>DCA</td>
<td>Custom</td>
<td>$\pi_j f_j(x_i</td>
</tr>
<tr>
<td>S</td>
<td>$y_i$</td>
<td>$\begin{cases} 1 &amp; y_i = j \ 0 &amp; y_i \neq j \end{cases}$ $\pi_j f_j(x_i</td>
</tr>
</tbody>
</table>

From a factor graph perspective [3, chap. 8.4.3], the basic premise of CA and WCA is the provision of additional (compared to US) information through the messages passed to the latent nodes $z_i$. As evident in the E-steps of Table 1, US only benefits from evidence $X$, while belief propagation with CA and WCA should be richer due to the additional contextual variables $C$. Of note, the original $S/US$ estimation problem $\theta$ is not cumbered with additional parameters related to the contextual variables, despite the model augmentation, due to the assumption of known priors and conditionals. It follows that the graphical representation of contextual assistance can be more complex than a single variable $c$, as long as the conditions of no additional parameters and seamless message parsing are satisfied. Essentially, the need for ground truth is replaced by a lesser requirement for knowledge of the aforementioned distributions. Those can be learned prior to the deployment of the algorithms, or even be publicly available (e.g., language models).

As already illustrated (Table 1), the context-related terms of each algorithm can be isolated to implicitly define sample-wise probabilistic “labels” $p_{ij}$ with $\sum_{j=1}^{M} p_{ij} = 1$ (i.e., each $p_i$ is a discrete probability distribution over the latent variable $z_i$). The entropy of these labels represents a measure of the contextual information content individually for each sample $x_i$ and, by averaging, for the overall estimation problem. Our work is the only one, besides [39], highlighting the importance of side-information measurability for the prediction of estimation benefits, our primary axis of investigation.

3.2. Information matrices and missing information principle

In order to shed light on the fundamental issue of alleviating the missing label information through the provision of side-information, one can rely on the Fisher Information [40], the most formal way of measuring the amount of information involved in the estimation of the unknown parameters $\theta$. Therefore, we study approximations of the (expected) Fisher information matrix $I(\theta)$ through its sampled-based version, the observed information matrix $I(\theta|X)$. The latter measures the amount of information a sample $X$ carries on the estimated
parameters $\theta$, where $I(\theta) = \mathbb{E}_\theta[I(\theta|X)]$ and $I(\theta|X) = -\frac{\partial \log L(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}_{ML}}$, the negative of the Hessian of the log-likelihood objective function evaluated at the ML estimate.

In [6], it is proved that the observed information for missing–data problems can be computed as the difference $I(\theta|X) = I_c(\theta|X) - I_m(\theta|X)$. The first term is an estimate of the available information if there were no missing data. The second term, called the missing information matrix, represents the information lost due to missing data. This relation has been called the missing information principle (MIP). Both these matrices can be computed through complete–data quantities (so that their calculation is tractable), as: $I_c(\theta|X) = \mathbb{E}_\theta\{-\frac{\partial^2 \log L_c}{\partial \theta \partial \theta^T}\} |_{\theta = \hat{\theta}_{ML}}$ and $I_m(\theta|X) = \text{cov}_\theta\{S_c(X|\theta)S_c(X|\theta)^T\} |_{\theta = \hat{\theta}_{ML}}$, where $S_c(X|\theta)$ is the score (gradient vector) of the complete–data log-likelihood.

The Fisher information also allows for the computation of the variance–covariance matrix of the MLE, as $C = I^{-1}(\theta|X)$ and, hence, the standard errors of parameter estimation as $SE_i = \sqrt{I_{i,i}^{-1}(\theta|X)}$ for the $i^{th}$ parameter in vector $\theta$, without resorting to repeated sampling. The same is true for the algorithms’ convergence rate, since, when EM converges to a local maximum, it has been shown [41] that the convergence rate $r = \lim_{k \to \infty} \|\hat{\theta}_k - \hat{\theta}_k^{+}\|$ is linear and coincides with the spectral radius ($\lambda_{max}$, where $\lambda_i \in [0,1], \forall i$, the eigenvalues) of the “rate” matrix $J$, defined as $J(\theta) = I_c^{-1}(X|\theta)I_m(X|\theta)$. The latter expresses the total fraction of missing information [42]. In Section 4 we use the definition $r' = 1 - r = 1 - \lambda_{max}$, which complies with the intuition that 0 corresponds to non-converging and 1 to immediately converging algorithms.

### 3.3. Evaluation metrics and simulation design

The results of Section 4 compare five algorithms ($CA$, $WCA$, $DCA$, $US$, $S$) in simulations with artificial datasets, as well as in real–world problems. The estimation properties reported are precision, standard errors and convergence rate. Estimation precision is the Euclidean distance between the estimated parameter vector $\hat{\theta}$ and the actual one $\theta^A$, namely: $D = \|\hat{\theta} - \theta^A\|$. For standard errors we employ the aforementioned estimator $SE_i$. For brevity, we
only report the average $ASE = \frac{1}{L} \sum_{i=1}^{L} SE_i$ (where $L$ is the number of estimated parameters). Similarly, the MIP–based estimate $r' = 1 - r$ is used for the convergence rate. The classification performance of trained models is assessed through N-class accuracy $A = N_c/N$, where $N_c$ the number of correctly classified samples out of $N$ total samples across all classes. Finally, for regression tasks, the mean square error $MSE$ is reported.

In order to quantify the information content of side–information, we employ a scaled negentropy definition on probabilistic labels $p_i$: 

$$NE_i = 1 + \sum_{j=1}^{M} p_{ij} \log M p_{ij}.$$ 

This metric is conveniently bounded, $NE_i \in [0, 1]$, for any number of mixtures $M$. $NE_i = 0$ when $p_i$ is uniform, $p_{ij} = 1/M, \forall j \in [1, M]$ (“ignorant” context, $p_i$ does not cast a preference over any class). Conversely, $NE_i = 1$ when $p_{im} = 1, m \in [1, M]$ and $p_{ij} = 0, \forall j \neq m, j \in [1, M]$ (“perfect” context, fully revealing the class label $y_i$. The $NE$ level of a dataset is extracted as the average across all included labels $p_i$.

For our simulation studies, a label $p_i$ for each sample $x_i$ is constructed randomly, so that its information content is $NE_i$. For all but one examined scenarios, all samples in $X$ are assigned the same $NE$ value ($NE_i = NE, \forall i$). In the “mixed” context scenario, however, each $NE_i$ is randomly drawn from a fixed interval. In all but the “wrong” context scenario (see below), $p_i$-s are constructed to cast greater confidence to the ground-truth label $y_i$ (“correct” context). Formally, we impose $\arg\max\{p_i\} = \arg\max\{y_i\}$, so that $p_i$-s always “predict” the correct $y_i$ with increasing confidence as $NE$ increases. This rule is only abandoned in the “wrong” context scenario, where the effects of misleading contextual information are investigated. In this scenario, $k_i = \arg\max\{p_i\} \neq \arg\max\{y_i\}, k_i \in [1, M]|i$ is selected randomly out of the $M - 1$ remaining possibilities for a reported percentage of the generated $p_i$-s.

The following scenarios are considered with “correct” context. A: a mixture of two univariate normal distributions, where variances are known and only the two class means are estimated, B: a mixture of two univariate normal distributions, where all existing parameters are estimated, C: a mixture of three univariate normal distributions, D: a mixture of two multivariate (2D) normal
distributions, E: a mixture of two univariate Maxwell–Boltzmann distributions and F: a mixture of two univariate, first order, linear regressors. These six scenarios are chosen to differ in terms of the numbers and types of mixtures employed (where, the Maxwell–Boltzmann of Scenario E is not a member of the exponential family), the number of estimated parameters, the dimension of the input space and the utility of the FMM (classification versus regression).

The scenarios targeting “mixed” and “wrong” context situations (Appendix F) employ mixtures of two univariate normal distributions.

For each scenario, 1000 estimation problems are generated and solved for all compared algorithms. Each problem $r \in [1, 1000]$ is associated to a randomly generated dataset $X_r, Y_r, P_r^{NE}$ of observed data, ground–truth labels and probabilistic labels of $NE$, respectively. For algorithms $CA$, $WCA$ and $DCA$, each problem $r$ is further solved for $NE \in [0 : 0.1 : 0.99]$, so that our evaluation encompasses the complete range of possible contextual information content. The cardinality $N$ of each dataset is fixed to 100 times the number of parameters to be estimated. The ground–truth $Y_r$ is constructed to have balanced number of samples per class. The observed data $X_r$ are randomly generated from semi-randomly selected “actual” distributions with parameters $\theta_r^A$ (of the respective scenario’s type) and the estimation begins with semi-randomly chosen initialization $\hat{\theta}_r^0$ (common to all algorithms). These semi-random procedures, detailed in Appendix C, ensure balanced number of samples per class and minimal impact of separability and initialization on the extracted results. All algorithms are left to perform as many iterations $t$, as needed so that $\| \hat{\theta}_r^t - \hat{\theta}_r^{t-1} \| < 10^{-5}$. If this stopping criterion is not reached after 300 iterations for some algorithm, $\hat{\theta}_r^{300}$ is used as its final estimate.

The classification accuracy $A$ is computed for each scenario, problem $r$ and algorithm, by generating a second “testing” dataset $X'_r, Y'_r$ (of equal cardinality to $X_r$) from the same “actual” FMM, which is classified using the estimated parameters of each algorithm by means of the Maximum–A–Posteriori rule. For the mixture–of–regressions scenario, the same evaluation methodology is applied to derive the $MSE$ on the testing set.
4. Results

4.1. Results on scenarios with artificial data

Appendix D justifies theoretically the estimation benefits brought forward by the proposed approach. More specifically, in Appendix D.1 we show, on the one hand, how the log-likelihood objectives in this methodology, compared to the \( US \) equivalent, exhibit lower local maxima except for the one closer to the supervised estimate. This effect increases the chances of convergence to this favourable extremum, while also reducing the EM algorithm’s sensitivity to initialization. On the other hand, Appendix D.2 shows that the fraction of missing information of an estimation problem, as expressed by the spectral radius of the rate matrix (see Section 3.2), is shown to decrease proportionally to the information content of the provided side–information, bounded by \( US \) and \( S \) (maximum and zero missing information, respectively). Therefore, the estimation properties that depend on the fraction of missing information, namely, the standard errors and the convergence rate, also benefit from context–awareness. Appendix E formally proves that these effects generalize to all FMMs. The theoretically anticipated effects are verified by simulations with artificially generated data, presented in this section.

The following set of simulations is meant to compare the performances of the derived algorithms for the “correct” context situation. Within each scenario, we illustrate each metric’s average across all 1000 problems for algorithm \( \alpha \) and some \( NE, \overline{M}_\alpha^{NE} \), normalized within the corresponding \( S \) and \( US \) performances, as: 
\[
\overline{M}_\alpha^{NE} = (\overline{M}_\alpha^{NE} - \overline{M}_{US})/(\overline{M}_S - \overline{M}_{US}).
\]

Fig. 2a–b show that, for all examined properties, \( CA \) and \( WCA \) exhibit improved performances proportionally to the strength of contextual assistance. Additionally, the average performances are upper–bounded by \( S \) (at \( NE \to 1 \)) and lower–bounded by \( US \) (at \( NE \to 0 \)). \( CA \) is consistently outperforming \( WCA \) for the same \( NE \), while both algorithms yield substantially better \( D, A \) and/or \( MSE \) than \( DCA \). While, as already commented, \( WCA \) reduces to \( US \) for \( NE = 0 \), \( CA \) yields improvements over \( US \) even in this case. Scenario A is the
Figure 2: Normalized, average (across 1000 problems of each of 6 FMM estimation scenarios A–F): (a) estimation precision $D$ (right y-axis only refers to DCA performances), (b) convergence rate $r'$, (c) average (across parameters) standard error $ASE$ and (d) classification accuracy $A$/regression mean square error $MSE$. The performances of each context–aware algorithm (CA, WCA, DCA) for each scenario A–F are color– and shape–coded as shown in the legends. Performances for different $NE$ levels arranged along the horizontal axis. For each metric and scenario, the embedded colormaps illustrate the lowest (when $US$ is involved) or highest (for all other combinations) $NE$ level (color–coded as shown in the adjacent colorbars) above which the algorithms on top and bottom of each column significantly differ at the 95% confidence interval (one–sided, paired Wilcoxon rank–sum tests with Bonferroni correction for multiple comparisons). Red color denotes no significant difference for any $NE$ level.

discussed exception with fixed $\pi$. These trends are universal, although the magnitude of improvements as a function of $NE$ is metric– and scenario–dependent, confirming the generalizability of the examples in Sections D.1 and D.2.

Statistical testing reveals that the added value of CA and WCA over conventional unsupervised estimation $US$ is significant for all metrics, already at very low $NE$ (for CA, even for ignorant context). Scenarios E and F for metrics $A$ and $MSE$, respectively, are exceptions where improved estimation precision $D$ does not translate into significantly better classification/regression, as an intrinsic property of the respective FMMs. Nevertheless, for metrics $D$ and $ASE$, it is
only at very high $NE$ that $CA$ and $WCA$ become indistinguishably similar to $S$, while for $r'$ supervised learning $S$ is significantly better even at $NE = 0.99$. The undesirable significance notwithstanding, the average differences for all metrics tend to operate much closer to the $S$ rather than the $US$ “extremity” even at low $NE$ (especially for $CA$). Furthermore, concerning $A$ and $MSE$, context–aware algorithms are statistically similar to $S$ since very low $NE$. The previously described and justified superiority of $CA$ over $WCA$ is shown to be significant only for the first few tested $NE$ levels for metrics $D$, $ASE$ and $A/MSE$, while it persists for almost the entirety of the $NE$ spectrum for $r'$. Finally, $DCA$ is again shown to suffer compromised performances.

Last but not least, our simulations demonstrate that context–awareness can substantially reduce the number of problems that could not converge with regular $US$ training\footnote{An EM algorithm fails to converge to a local maximum when the spectral radius of $J$ exceeds unity [42].}. In the most characteristic example (due to the larger number of mixtures) of scenario C, where 83.7\% of problems did not converge with $US$, this percentage is reduced by $CA$ to 3.8\% at $NE = 0$ and 1.0\% for $NE \in [0.1, 0.99]$. $WCA$ also alleviates this problem (less aggressively), by gradually reducing the non-convergence percentage to 67\% at $NE = 0.1$, 39.4\% at $NE = 0.5$ and, eventually, 1.0\% at $NE = 0.99$. It is thus shown that context–awareness is able to avoid irregularities, already at low $NE$ levels. In the interest of space, results on more “realistic” situations where contextual assistance can be, to some extent, “wrong” and of “mixed” $NE$ are offered in Appendix F.

4.2. Results on real–world scenarios

4.2.1. Online learning in brain–computer interface

The applicability and effectiveness of context–aware learning are demonstrated in an online–learning problem of brain–computer interaction (BCI). More specifically, we employ the binary “BrainTree” speller described in [43], where a 2-class, motor imagery (MI) BCI translates electroencephalographic
(EEG) brain patterns into one of two required control commands.

![Graphical user interface of the BrainTree speller (top) and its underlying binary tree structure (bottom).](image)

**Figure 3:** (a) Graphical user interface of the BrainTree speller (top) and its underlying binary tree structure (bottom). (b) Mean and standard deviation of balanced running classification accuracy during the spelling tasks for each of the 12 subjects with supervised estimation $S$, as well as algorithms $CA$ and $CAE$. Red asterisks on top of $CA$ bars denote statistical significant difference with $CAE$ (Wilcoxon ranksum test, $\alpha = 0.01$). No statistical significant differences between $CA$ and $S$ are found. The last triplet of bars illustrates the averages across subjects.

BCI is known to suffer from non-stationarity of brain patterns, what degrades previously trained classifiers and calls for online classifier learning [44]. Yet, the latter has to be carried out in an unsupervised manner, since data labels cannot be retrieved during online BCI operation. Consequently, online, adaptive classifier training is bound to suffer the known shortcomings of unsupervised learning. However, since the speller provides a natural candidate for extraction of contextual assistance, it can be expected that context–aware learning could allow uninterrupted BCI spelling.

Fig. 3a (top) illustrates the speller’s graphical user interface (GUI), where characters are arranged alphabetically. A vertical red cursor, the “caret”, denotes the current position, while an orange “bubble” surrounds the characters currently available. Underneath the character bar, the user observes a conventional MI BCI feedback. The user employs one of two MI tasks to move the caret towards the desired character. The procedure is repeated until the latter is the only one left within the “bubble”, in which case it will be typed after the next transition and a new typing round is initiated. This simple GUI hides the underlying complexity, where characters are the leaf nodes of a binary tree.

A simple example on a reduced dictionary is illustrated in Fig. 3a (bottom).
Therein, the caret represents the tree’s current internal node, the “bubble” surrounds the leaf node characters within the current node’s two subtrees and a BCI command moves to the left/right child node. Effectively, in each typing round, each character is associated with a binary “codeword” of left/right transitions.

This structure provides a straightforward mechanism for retrieving contextual assistance through the speller. By modeling the current (at time $t$) desired character as a contextual random variable $c_t \in \{a,b,\ldots \text{backspace}\}$, according to the definitions for CA (Table 1), one only needs to know the priors $p(c_t)$ and conditionals $p(z_t|c_t)$ (where $z_t \in [0,1]$, the MI class the user is currently employing). CA learning with unobserved context is applied, since the speller is unaware of the user’s desired character. A trained Prediction by Partial Matching (PPM) language model provides the priors $p(c_t)$ for each typing round, based on the currently written prefix. Conditionals $p(z_t|c_t)$ are also easily extracted given the structure of the tree and knowledge on the current node position.

More specifically, $p(z_t = j | c_t = k) = 1$ holds if the character $k$ is a member of the current subtree $j \in \{\text{left, right}\}$ and $p(z_t = j | c_t = k) = 0$, otherwise.

A custom type of binary tree able to provide $p_t$-s of high $NE_t$ is employed. Each node’s subtrees are arranged to obey as much as possible a 0.9/0.1 or 0.1/0.9 split of total character probability (the “heavy” subtree is reversed at each level to avoid class–bias), while still maintaining alphabetic ordering. This results in a “mixed” context scenario, since the aforementioned split is not always possible given any position in the tree and the current $p(c_t)$. Small percentages of “wrong” context also exist, as in some cases the desired character is not a member of the “heavy” subtree.

We devise a buffer approach for continuous, context–aware, online–learning of a BCI classifier, modeled as a mixture of two multivariate, 6–dimensional normal distributions with common covariance matrix (effectively, an LDA classifier). Six features capturing a subject’s spatially distributed sensorimotor rhythms are extracted in a sliding window, twice per second (2 Hz). EM–learning takes place in a buffer of the latest two minutes of data (240 feature vectors/samples). Consecutive buffers are shifted by only 1 sample, thus a new,
slightly updated classifier is used to classify each incoming sample. The features are log-transformed and known to be approximately normal.

We conduct spelling simulations using EEG MI data of 12 subjects recorded with the 2-class MI BCI protocol described in [45]. For each evaluated algorithm and subject, data are “played-back” in the order recorded. A common subject-unspecific classifier is used as the initial point of adaptation. The spelling task consists of the words “nothing” and “portion”, and is repeated for algorithms $S$, $CA$, as well as $CA$ with ignorant context (noted $CAE, NE_t = 0, \forall t$). Automatic correction of erroneous trials is imposed, simulating the correction mechanism users employed in [43]. The corresponding online experiment has been also successfully conducted [46]. $CA$ is expected to outperform $CAE$ in terms of classification accuracy, not only due to its superior estimation precision, but also, thanks to its improved convergence rate. That is because a 100 msec/sample time limit for all processing is imposed to cope with the application’s real-time demands, so that $CA$ has higher chances of convergence within the limited amount of iterations executed. $CAE$ replaces $US$, since the latter yields nearly chance-level accuracies for all subjects, as a large number of parameters needs to be learned from only 240 samples in each buffer and only a few iterations. We calculate “balanced” (average of class-recalls), “running”, 2-class classification accuracy $BA$ in a window of the latest minute (120 samples) of simulated BCI spelling, with a shift of 30 seconds (60 samples), a sort of prequential evaluation akin to online learning.

Fig. 3b shows the average and standard deviation of running $BA$ for each subject (as well as the average across subjects) after the first minute of spelling. $CA$ learning is shown to yield similar $BA$ to $S$. That holds for each subject individually, as well as for the average across subjects. It is further considerably better than the $CAE$ algorithm (which is close to chance level for most subjects), demonstrating the importance of contextual assistance. The case of subject $s5$, where all algorithms perform poorly, was found to be due to particularly intense instabilities of the brain activity. Furthermore, $CAE$ for subject $s3$ demonstrates a class-inversion effect because, although this subject has gen-
erated discriminant brain patterns, CAE cannot escape the local maxima near the initialization point. On the contrary, for s8, initialization is coincidentally favorable, thus CAE is only slightly inferior to CA and S.

Overall, this exemplary application showcases the possibility to intuitively apply context–aware parameter estimation in real–world problems. Context awareness accounts for performances approaching those of supervised learning, despite any manual collection of ground truth is redundant. Improvements over unsupervised learning are achieved despite only “mixed” and “wrong” context is available. The added value for BCI is that users can communicate “on–demand”, without resorting to conventionally used, but, lengthy and cumbersome supervised retraining. The latter interrupts normal operation and limits deployment of BCI in every–day life of people with disabilities.

4.2.2. Unsupervised detection of breast cancer

The BCI example illustrates the applicability of the CA algorithm with latent context. We additionally demonstrate the application of CA with observed context and of WCA in another real-world scenario concerning the detection of malignancies in mammograms, a case of general computer-aided diagnosis (CAD). CAD systems are pattern recognition devices exploiting diagnostic factors extracted from medical exams, such as biomedical imaging and biochemical tests, which form the features/predictors (variable $x$ in the present formalism) of the system, in order to make improved diagnostic decisions (i.e., identify the underlying medical condition $z$).

There are several reasons rendering CAD an ideal usage scenario for the examined algorithms. First, the difficulty and cost of collecting large databases for CAD favor simple models for decision making to avoid overfitting, like the FMMs exemplified here. Second, exact inferences based on medical tests to provide labeled instances can be laborious, painful (e.g., biopsy), expensive and even life-threatening (e.g., amniocentesis to verify other prenatal tests). As a result, there are nowadays many “unlabeled” medical data that cannot be useful in conventional supervised CAD, implying a great need for unsupervised
learning in this area.

Importantly, CAD scenarios are very likely to enjoy an abundance of exploitable side–information to improve unsupervised learning for both CA and WCA: On the one hand, known or easily collected risk factors like age, family history, genetic and environmental indicators (e.g. smoking for lung cancer) can play the role of “contextual” factors \( c \) implicated in the appearance of a disease \( z \), as in Figure 1a, giving rise to the applicability of CA with observed context. On the other hand, additional diagnostic factors depending on the existence and type of the disease, but conditionally independent from the medical test’s features \( x \), as in Figure 1b, yield the possibility of WCA learning. Besides conditional independence \( x \perp\!\!\!\!\!\!\!\perp c \mid z \), different set supports (e.g. real versus categorical) or different distributions \( p(x \mid z) \) and \( p(c \mid z) \) might prevent embedding \( c \) as an extra feature into \( x \), thus dictating the use of WCA.

![Figure 4](image-url)

**Figure 4:** (a) Example of segmented breast tumor image from the CBIS-DDSM database. (b) Mean and standard deviation (across 1000 repetitions) of testing set classification accuracy for benign and malignant mammogram tumors with LDA classifiers learned through algorithms \( S \), \( WCA \), \( CA \) with observed side–information and \( US \). All illustrated accuracy distribution differences are statistically significant after Bonferroni correction (paired, two–sided Wilcoxon signed–rank tests, \( \alpha = 0.01 \)).

We employ 511 (261 benign vs 250 malignant) segmented (circumscribed or spiculated) breast tumor image samples (see example in Figure 4a) from the publicly available and pathologically verified CBIS-DDSM mammography database [47]. Five tumor shape and texture descriptors identified as optimal in [48] form our feature set and LDA classification is used for this exemplary
CAD. We perform 1000 repetitions randomly shuffling the dataset, keeping 60% of 500 samples as the training set for estimating the classifier’s parameters (class-dependent mean vectors and common covariance matrix) and testing on the remaining 40% samples. For the unsupervised EM algorithms WCA, CA and US, 11 samples (6 benign and 5 malignant) are used to produce a more precise initial estimate than that achieved by random initialization, simulating the case where the pathology of at least a few samples has been verified with biopsy (exact labeling). All algorithms are left to perform 100 iterations for each testing repetition. For the WCA algorithm, side–information $c$ consists in the BI-RADS assessments (manual interpretation of mammograms) offered in the CBIS-DDSM metadata. The WCA statistics $p(c), p(c|z)$ are estimated from all database samples. For compact, comparative demonstration of CA with observed context, in the absence of any risk factor information in the database, such a factor is artificially constructed. Specifically, we assume the existence of a binary risk factor$^4$ $c \in \{\text{present, absent}\}$, present in 50% of the population $p(c = \text{present}) = 0.5$, which is highly indicative of malignancies: $p(z = \text{Malignant}|c = \text{present}) = 0.7$, $p(z = \text{Malignant}|c = \text{absent}) = 0.2$. We randomly distribute contextual values $c_i$ across our 511 samples according to these priors and conditionals for each testing repetition.

Figure 4b verifies that a breast cancer CAD system trained in the absence of ground truth, but enjoying contextual assistance, is able to significantly outperform a CAD trained in a classical unsupervised manner and approach the performances of $S$. Additionally, it is evident that the detection performance stability is also improved, reflected in reduced standard deviations across repetitions for the context-aware algorithms. As anticipated, classification improvements follow those of the estimation properties (precision, standard errors, convergence rate), using the equivalent estimates of $S$ as measure (data not shown).

$^4$Such a factor could be the existence or not of mutations in the BRCA1 and BRCA2 genes.
4.2.3. Unsupervised recalibration of fingerspelling classifiers

Exploiting contextual information is further exemplified with a fingerspelling application. The goal is to automatically recognize from video frame sequences, sign language hand shapes/gestures, where each sign corresponds to a character. Such a device allows non-vocal communication and control, particularly useful for domotics and accessibility of people with speech disorders.

We make use of the publicly available American Sign Language (ASL) dataset described in [49], which consists of a Microsoft Kinect RGB and a depth video frame for each gesture. The dataset includes more than 500 signs for each of 24 characters (j and z excluded as their corresponding gestures involve motion) and 5 different subjects, for a total of more than 60000 available samples. For each sample, a 1000-dimensional representative feature vector is extracted by applying Gabor filtering on both RGB and depth images and concatenating the result, as proposed in [49].

A plug-and-play fingerspelling recognition system comes with a pre-trained classifier (conventionally, subject-unspecific, i.e. based on multi-subject datasets), that is expected to adequately generalize to a new user, so that a cumbersome user-specific calibration imposing copy-spelling tasks can be avoided and the client can immediately use the system for free spelling. Nevertheless, such classifiers are known to underperform compared to those trained on one’s own data, as the latter are able to better capture user- and ambient-specific features. Starting with the subject-unspecific classifier, unsupervised learning can be employed to recalibrate the classifier accordingly, so as to improve system performance. The latter procedure could benefit from the unsupervised context-aware parameter estimation methodologies proposed in this work.

Specifically, similarly to the exemplary application of Section 4.2.1, we take advantage of the rich side-information embedded in language. In this case, individual words are easily segmented thanks to the naturally occurring pause in-between consecutive words. As shown in the graphical model of Figure 5a, one observes the (potentially error-inflicted) $k^{th}$ written word $o_k$, corresponding
to (and depending on) the user’s intended word \( w_k \). Each of the \( L \) consisting characters of \( w_k \) at the \( i^{th} \) position, \( z_i \), also naturally depends on the desired word \( w_k \) and gives rise to the observed feature vector of a sign \( x_i \). Of note, the characters \( z_i \) are also the latent class labels of the classification problem. Ignoring for simplicity the dependence among characters, the joint probability distribution of a single character \( z_i \) according to this graphical model is:

\[
p(x_i, z_i, w_k, o_k) = p(x_i|z_i)p(z_i|w_k)p(o_k|w_k)p(w_k)
\]

so that the conditional distribution given observed \( x_i, o_k \) is:

\[
p(z_i, w_k|x_i, o_k) = \frac{p(x_i|z_i)p(z_i|w_k)p(o_k|w_k)p(w_k)}{p(x_i)p(o_k)}
\]

where the term in the denominator represents a normalization factor \( K \). Marginalizing out the latent variable \( w_k \) yields:

\[
p(z_i|x_i, o_k) = \frac{p(x_i|z_i)}{K} \sum_{w_k} p(z_i|w_k)p(o_k|w_k)p(w_k)
\]

The forms of Fig. 5a and Equation 4 imply CA type of learning (see Table 1), with probabilistic labels:

\[
p_{ij} = \sum_{w_k} p(z_i = j|w_k)p(o_k|w_k)p(w_k), j \in \{a, b, \ldots y\}
\]

The terms appearing in this definition are estimated as follows: The word priors \( p(w_k) \) are either calculated by enumerating the word occurrences in a text corpus, or, alternatively, a uniform word distribution may be assumed for simplicity. The conditionals \( p(z_i = j|w_k) \) can only take on value 1, if the word \( w_k \) contains character \( j \) at position \( i \), or 0 otherwise. The term \( p(o_k|w_k) \) expresses the probability that the written string \( o_k \) could be the result of the user attempting to spell the word \( w_k \). It can be therefore approximated through any kind of string similarity distance. For this work, the popular and simple Levenshtein distance is employed.

The existence of this term reveals a certain sensitivity of this type of contextual assistance on the initial, subject-unspecific classifier. Considerably compromised classification accuracy yields many erroneously spelled characters in
\( o_k \), which results in low probability \( p(o_k|w_k) \) even for the ground truth \( w_k \). In other words, the ability of the observed word \( o_k \) to adequately reveal the real underlying word \( w_k \) will be limited. In addition to this, the summation over the possible words \( w_k \) in a text corpus implies that smaller corpora will yield richer contextual assistance, even if the possibility of accurate word segmentation allows to exclude from summation all words \( w_k \) whose length is different than that of the written string \( o_k \). That is, since an observed \( o_k \) will have less eligible candidates \( w_k \) to be matched to.

We assess the added value of \( CA \) in this application scenario by simulating spelling sessions for all 5 subjects \( A - E \) and 6 different text corpora of increasing size. For each subject, simulated spelling is performed with an initial subject-unspecific GMM classifier (with common covariance matrix for all classes) trained on the data of the remaining 4 subjects. Subsequently, a target word from the text corpus is randomly selected, representing the intended word to be spelled, \( w_k \). For each consisting character of \( w_k \), a representative sign/sample \( x_i \) from the subject’s database is randomly chosen and classified, producing the “written” word \( o_k \). A probabilistic label for each sample \( p_i \) is then constructed as described above. The procedure is repeated until all the subject’s available samples have been used. We finally evaluate the recalibration effectiveness of different methods through 5-fold cross-validation.

The methods tested are: i) using the initial subject-unspecific classifier (\( SU \)), where the training folds are not used at all, and classification accuracy is computed on the testing fold of each cross-validation iteration, ii) supervised recalibration (\( S \)), where new classifiers are learned on the training folds using the ground truth labels (as could be derived with a copy-spelling user training procedure), iii) unsupervised recalibration (\( US \), where the standard EM algorithm is applied on the training folds, initialized with the subject-unspecific classifier and iv) context-aware unsupervised recalibration (\( CA \). Iterative algorithms \( CA \) and \( US \) are left to perform 5 iterations, which is shown to be enough for convergence. Furthermore, only the class mean vectors are re-estimated, which has been found to avoid overfitting and, as anticipated, decrease the computational
complexity.

Since the strength of this type of context depends on the number of words that can be used, we test the CA algorithm on 6 different text corpora of size \( N = 20, 100, 250, 500, 1000 \) and 10000. The smallest corpus (CA-20) includes twenty words useful as commands for a smart house control application (e.g., “lights on/off”). The biggest corpus CA-10000 is an entire novel comprising approximately 10000 different words. The remaining corpora used, CA-N, consist of the \( N \) most common English words. Given the anticipated superior performance of CA recalibration for small allowed word sets, the user is supposed to select his own preferred point with respect to the trade-off between decoding performance and the number of words he is allowed to use. It is also straightforward to implement a strategy where a new user is at first limited to small corpora, until the quality of recalibrated classifiers improves and he/she can gradually move to bigger corpora and, eventually, completely free spelling.

This application substantiates that the proposed method can be effective even in highly multivariate (1000 features) and multi-class (24 classes) problems. Additionally, it confirms that contextual variables may form more complex networks than the simple cases of Fig. 1, with no implications whatsoever on the algorithms’ properties and effectiveness. Furthermore, as in the previous scenarios, it is shown that the proposed approach can deliver benefits despite the retrieved contextual assistance may have shortcomings. Specifically, for the 6 corpus sizes tested, the strength of contextual information \( NE \) drops proportionally (0.90, 0.64, 0.50, 0.44, 0.35, 0.22), as does the amount of “correct” context, quantified as the percentage of samples where the probability of the ground truth class exceeds 0.5 (i.e. dominates the respective probabilistic label \( p_i \)): 90.2, 41.5, 28.6, 21.2, 10.9, 5.0%.

Figure 5 illustrates the results of the spelling simulations. First, it is evident that in this scenario subject-specific calibration is indeed necessary, as in all cases (except for subject D) even the naive unsupervised method US (red) outperforms the subject-unspecific classifier (white). Most importantly, context-aware adaptation is superior to the conventional approach US, on average by
3.5% \((p = 0.035\) with unpaired, two-sided Wilcoxon ranksum test\) even at the weakest level of contextual assistance \(CA-10000\) (darkest blue). The classification accuracy gains are inversely proportional to the size of the dictionary used (shades of blue). When applying the method that maximizes accuracy, \(CA-20\), the average improvement over \(US\) reaches 19.5% \((p < 10^{14})\). This performance is still 9.6% below that of the supervised method \(S\) \((p < 10^{14})\). Overall, in the most demanding application scenario tested, the conclusions remain consistent: \(CA\) parameter estimation yields classification accuracy improvements proportionally to the information content of the provided contextual assistance, and its performance is bounded by the equivalent performances of \(S\) (upper) and \(US\) (lower) on the same problem.

5. Discussion

This work has studied unsupervised MLE algorithms devoid of any need for data labels, but able to exploit side-information in the form of probabilistic

![Figure 5: Context-aware learning in a fingerspelling application scenario. (a) Probabilistic graph of contextually enhanced FMM. (b) Average cross-validation classification accuracy of FMM classifiers trained with different algorithms as color-coded in the legend.](image-url)
context embedded into a generative model and with known statistics. A com-
parative analysis and in–depth study of these algorithms’ properties for finite
mixture models is offered from both a theoretical and a practical standpoint.

It can be argued that from the modeling and mathematical derivation view-
point, these algorithms hardly qualify as a novel methodology for unsupervised
learning (straightforward application of EM on slightly augmented graphs).
However, it is clear that the literature of learning from side–information has
ignored the suitability of such fundamental techniques in favour of less practical
alternatives. As an example, we find that in most application scenarios and for
most prospective users, it should be possible and much more intuitive to ex-
press a given type of side–information as a random variable with known statis-
tics rather than through a generalized expectation criterion. In other words,
the admittedly greater generality and flexibility of the latter can be most often
traded off (Appendix A). Simplicity of the resulting derivations and formula-
tions can only be viewed as additional advantages. We thus consider bringing
this methodology in the spotlight to be the main contribution of this article.

A second unique contribution entails the identification of basic principles
giving rise to improved EM–MLE by context–awareness. First, we have shown
that a context–assisted log–likelihood objective is favourably distorted in com-
parison to the regular one, so that sensitivity to initialization diminishes and
the chances of convergence closer to the supervised MLE increase. The second
principle regards the partial elimination of missing label information through
context as a result of the applicability of the MIP. Demonstrating this makes
our work the first one to justify the benefits of side–information in learning
from an information–theoretic viewpoint. Through these principles, we have
established experimentally and, wherever possible, also formally, two impor-
tant points. First, that any positive effects on the estimation properties are
proportional to the information content of implicitly extracted instance–wise
probabilistic labels. Second, that the proposed algorithms perform between the
boundaries defined by the unsupervised and supervised equivalents of a given
problem. Future work could extend this line of research in various directions,
e.g., establishing a link between side-information and estimation precision.

We have showcased that estimation benefits are still evident and significant in problems with variable, weak, or even, to a certain extent, “wrong” contextual assistance, situations likely to arise in practical applications. Furthermore, the algorithms’ limitations and comparative advantages have been outlined. In this regard, we have demonstrated that completely disregarding the evidence from observed samples in favour of context, like with the DCA algorithm, yields inferior estimation properties in spite of immediate convergence. A general superiority of CA over WCA as a result of removing missing information related to the mixing coefficients has also been demonstrated. Furthermore, the application of all algorithms in tough, real-world problems, showcases the broad applicability of context-aware learning as proposed here.

As argued in Section 3, the main limitation of the algorithms proposed here is their non-universal applicability. However, this is not specific to the proposed framework, but, rather, a limitation shared among all methods exploiting side-information. Indeed, it is not guaranteed that for any application exploitable context as shown here exists, or that the cost of automatically retrieving contextual assistance will be lower than that of explicitly labeling data. However, as the real-world examples of this paper illustrate, rich context should be easily and cheaply acquired in a broad application spectrum.

Another limitation regards the fact that the prerequisite knowledge of distributions $p(z|c)$ or $p(c|z)$ implies that the latent class labels are at least defined, i.e., the number and type of mixtures/classes $M$ is known. Consequently, the proposed methodology regards unsupervised classification and not general clustering problems.

The scope of this work has been deliberately limited to FMMs, since it is the simplest and most popular generative model. Proving the generalization of our claims to all types of FMMs already accounts for extensive applicability of the presented algorithms. FMMs are in themselves a very general tool, by virtue of the possibility to replace the mixture types, number of mixtures, etc., to the ones suited to a particular problem. The demonstrated applicability
of our algorithms to the mixture–of–regressions scenario supports this claim. However, the outlined derivations prove that similar benefits can be derived for more complex Bayesian networks, like Hidden Markov Models.

Concluding, our work has established that the concept of context–awareness can play a key role in model learning. Future work could entail investigations of the effects context–aware learning has in Bayesian estimation and how our conclusions generalize for models other than FMM.

Acknowledgements

This work was supported by the European ICT Programme Project FP7-224631 and the Hasler Foundation, Switzerland. This paper only reflects the authors views and funding agencies are not liable for any use that may be made of the information contained herein.

References


Author biographies

**Serafeim Perdikis** received the M.E. degree in Electrical and Computer Engineering from the Aristotle University of Thessaloniki (AUTH) in 2007 and the Ph.D. degree in brain-computer interaction (BCI) from École Polytechnique Fédérale de Lausanne (EPFL) in 2014. In 2015 he joined the Wyss Center for Bio- and Neuro-engineering, Geneva. Since November 2017 he has been with Mindmaze SA as a research engineer and with EPFL as a post-doctoral fellow.

**Robert Leeb** was scientist and chief-engineer at the Center for Neuroprosthetics at École Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland, where he worked on the transfer of brain-computer interface (BCI) technology towards its clinical partner at the Clinique Romande de Réadaptation (SUVA). Furthermore, he was senior post-doctoral researcher at the Chair in Non-Invasive Brain-Machine Interface at EPFL, where he worked on hybrid BCI and the application of BCI technology towards patients. Currently, he works at MindMaze SA, Switzerland, where he heads the bio-signal activities for rehabilitation and gaming. He holds a M.Sc. degree in electrical and biomedical engineering from Graz University of Technology, Austria, and received his Ph.D. in Computer Science from Graz University of Technology, Austria.

**Ricardo Chavarriaga** is a senior researcher at the Chair in Brain-machine interface, Center for Neuroprosthetics, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. He holds an engineering degree in electronics from the Pontificia Universidad Javeriana (Cali, Colombia) in 1998, and a Ph.D. in Computational Neuroscience from the EPFL in 2005. His work is focused on the analysis of brain electrical signals and the design of brain computer interfaces.

**José del R. Millán** joined the École Polytechnique Fédérale de Lausanne (EPFL) in 2009 to help establish the Center for Neuroprosthetics. He holds the Defitech Foundation Chair and directs the Brain-Machine Interface Laboratory. He received a PhD in computer science from the Technical University of Catalonia, Barcelona, in 1992. Previously, he was a research scientist at the Joint Research Centre of the European Commission in Ispra (Italy) and a senior re-
searcher at the Idiap Research Institute in Martigny (Switzerland). He has also been a visiting scholar at the Universities of Berkeley and Stanford as well as at the International Computer Science Institute in Berkeley. Dr Millán has made several seminal contributions to the field of brain-computer interfaces (BCI), especially based on electroencephalogram (EEG) signals. Most of his achievements revolve around the design of brain-controlled robots. He has received several recognitions for these seminal and pioneering achievements, notably the IEEE-SMC Nobert Wiener Award in 2011 and elevation to IEEE Fellow in 2017. During the last years Dr Millán is prioritizing the translation of BCI to end-users suffering from motor disabilities. As an example of this endeavour, his team won the first Cybathlon BCI race in October 2016. In parallel, he is designing BCI technology to offer new interaction modalities for able-bodied people.
PrepNet: A Convolutional Auto-Encoder to Homogenize CT Scans for Cross-Dataset Medical Image Analysis

Mohammadreza Amirian*,†, Javier A. Montoya-Zegarra*, Jonathan Gruss*, Yves D. Stebler*, Ahmet Selman Bozkir*, Marco Calandri†, Friedhelm Schwenker† and Thilo Stadelmann†

*ZHAW School of Engineering, 8400 Winterthur, Switzerland
†Ulm University, Institute of Neural Information Processing, 89081 Ulm, Germany
‡University of Turin, Department of Oncology, 10124 Turin, Italy
§Fellow, ECLT European Centre for Living Technology, 30123 Venice, Italy
These authors contributed equally to this work

amir@zhaw.ch

Abstract—With the spread of COVID-19 over the world, the need arose for fast and precise automatic triage mechanisms to decelerate the spread of the disease by reducing human efforts e.g. for image-based diagnosis. Although the literature has shown promising efforts in this direction, reported results do not consider the variability of CT scans acquired under varying circumstances, thus rendering resulting models unfit for use on data acquired using e.g. different scanner technologies. While COVID-19 diagnosis can now be done efficiently using PCR tests, this use case exemplifies the need for a methodology to overcome data variability issues in order to make medical image analysis models more widely applicable. In this paper, we explicitly address the variability issue using the example of COVID-19 diagnosis and propose a novel generative approach that aims at erasing the differences induced by e.g. the imaging technology while simultaneously introducing minimal changes to the CT scans through leveraging the idea of deep auto-encoders. The proposed prepossessing architecture (PrepNet) i) is jointly trained on multiple CT scan datasets and (ii) is capable of extracting improved discriminative features for improved diagnosis. Experimental results on three public datasets (SARS-COVID-2, UCSD COVID-CT, MosMed) show that our model improves cross-dataset generalization by up to 11.84 percentage points despite a minor drop in within dataset performance.

Index Terms—Adaptive preprocessing, domain adaptation, auto-encoder

I. INTRODUCTION

A major challenge in rolling out machine learned models to a broad user base is the variability of data encountered in the real world. Models can only be expected to work well on data of similar distribution as has been used for training, but ubiquitously, differences in image acquisition setup hinder the applicability of a once developed model in novel settings. A recent example for the negative effects of such failure to adapt between different domains has been given at the start of the COVID-19 pandemic:

As of 2nd February 2021, this disease has caused over 100 million infections worldwide and over 2 million deaths according to the World Health Organisation (WHO) [1]. To alleviate this, rapid diagnosis of COVID-19 cases has been proven to be effective for decelerating the spread of the disease [2]. According to [2], [3], reverse transcriptase quantitative polymerase chain reaction (RT-qPCR) tests are accepted as the gold standard for the identification of positive cases. However, this type of test was not available in sufficient numbers at the beginning of the pandemic. Further, beyond being time-consuming, it relies on both human effort and expert knowledge. Thus, there arose a need for automatic diagnostic methods that can assist experts and reduce human efforts by targeting the automatic identification of COVID-19 positive cases. The literature has shown promising efforts in the automatic identification of COVID-19 cases from lung computed tomography (CT) scans using computer vision methods [4], [5], [6], [7]. Lessmann et al. addressed cross-vendor analysis (between different CT scanners such as Varian, Siemens, GE Healthcare, Philips and Canon) for 3D CT scans successfully [8]. However, it is demonstrated that a considerable drop in cross-dataset performance appears for the diagnosis of 2D CT scans acquired via different devices. Thus, the previously mentioned within dataset variability has the potential to discourage the community to merge and annotate data from multiple sources. As a result, combining datasets is a challenge posed not only for COVID detection but also for other applications in diagnosis and segmentation.

In this paper, we address domain adaptation of medical image analysis methods by proposing a deep convolutional neural network (CNN) for preprocessing 2D CT scans such that it is trained to fool a classifier that discriminates between various CT datasets, thus aiming to remove the within dataset variability. We evaluate the performance of the suggested method on the exemplary use case of predicting COVID-19 positive cases, due to the global variability in respective datasets and the availability of plenty of opportunities to compare. It should be noted that, the methodology is inspired by generative adversarial learning [9], [10]. Our contribution is twofold: (i) we propose a novel trainable preprocessing CNN for CT scans through leveraging the idea of deep auto-encoders.
architecture with a dual training objective that is capable of equalizing the variability of different CT-scanner technologies in the image domain as a pre-processor (PrepNet); (ii) we validate this model by showing the transferability of its diagnostic capabilities between different CT data sources based on common public benchmarks. We conduct experiments on the SARS-CoV-2 CT-scan dataset [11] and the UCSD COVID-CT dataset [12] as well as MosMed dataset [13]. Our results show that our PrepNet model improves the cross-dataset COVID-19 diagnosis performance (i.e., training on one dataset and testing on another) by 11.84 percentage points (pp) through creating a unified representation of multi-dataset CT scans.

II. RELATED WORK

With the emergence of COVID-19, many studies and datasets have been proposed in the literature which show an increase in data diversity over time and the extent of related computer vision methods to deal with it [14], [15]. Horry et al. [2] utilize a transfer learning scheme to build various COVID-19 classifiers based on several off the shelf CNN models such as VGG16/19 [16], Resnet50 [17], InceptionV3 [18], Xception [19], and InceptionResnet [20]. They compared the generalization capability of various images sources such as X-ray, CT and ultrasound images and developed a pre-processing scheme for X-ray images to reduce noise at non-lung areas in order to decrease the effect of quality imbalance among the employed images. A VGG19 [16] coupled with ultrasound images is found to yield the best validation accuracy of 99%, while 84% have been achieved using CT scans [21].

He et al. [21] propose a sample-efficient learning concept called “Self-Trans” via synergetically combining transfer learning and contrastive self-supervised learning. They seek intrinsic visual patterns in CT scans without relying on labels created with human effort. Besides, they open-sourced their CT dataset involving 349 COVID-19 positive patients and 397 COVID-19 negatives [12]. They achieve an accuracy of 86% through unbiased feature representations together with a reduction of overfitting.

Mobiny et al. [22] propose the DECAPS approach with following contributions: (i) inverted dynamic routing [23] to avoid seeking visual features from non-related regions, (ii) training with a two-stage patch crop and drop strategy to encourage the network to focus on the useful areas, (iii) employing conditional generative adversarial networks for data augmentation. Experiments result employing a pre-processing network (PrepNet) to standardize CT images with respect to the visual differences among datasets prior to training of any final diagnosis model, relying on generative architectures since they showed very promising results for similar tasks [22]. An advantage of this approach is that the PrepNet can be combined with any downstream diagnosis model, thus leveraging future progress there without additional costs while improving cross-dataset performance.

Two research papers closely related to the goal of domain adaptation in this study are presented by Lessmann et al. addressing cross-vendor diagnosis [8] and Amyar et al. using auto-encoders in multi-task learning [30]. Nevertheless, Lessmann et al. did not confront a considerable cross-vendor performance drop because of using a richer source of information (3D scans) as explained in [31]. Amyar et al. leveraged multi-task learning and trained an auto-encoder besides a segmentation and classification model for COVID-19 diagnosis. However, they did not aim at removing the cross-dataset variability of the scans. This study focuses on homogenizing the 2D CT scans by reducing cross-dataset information.

III. METHODOLOGY

In this section, we give details of our PrepNet model in terms of network architecture, core modules, and loss functions. The architecture of our proposed model is presented in Figure 1. For a group of $\mathcal{N}$ input CT scans $\{X^n\}_{n=1}^N$ coming from different CT vendors’ devices, our model extracts multi-scale discriminative feature maps through an auto-encoder and reconstructs the original CT scans $\{\hat{X}^n\}_{n=1}^N$. The reconstructed CT scans are next fed into a dataset/technology
classification branch which acts as a pseudo-label classifier and is responsible for discriminating among different CT datasets. Once this model is trained end-to-end in an adversarial way, the reconstructed CT scans are fed into a COVID-19 classifier which is trained directly on the reconstructed CT-scans. The COVID-19 classification branch is responsible for the classification of healthy vs. non-healthy patients. The complete network model with its main modules are described in more detail below.

A. Model Architecture

Auto-Encoder Module: We feed a CT scan image $X^n$ into our auto-encoder ($E_a$ and $D_a$) and obtain a reconstructed version $\hat{X}^n$ given by $\hat{X}^n = D_a(E_a(X^n))$. The encoder $E_a$ is based on the standard classification network VGG-Net [16], whilst the decoder $D_a$ is a convolutional network with the same number of layers as the encoder. We add skip-connections from $E_a$ to $D_a$ to recover the spatial information lost during the down-sampling operations.

Dataset Classifier Module: The CT dataset classifier $E_t$ receives the reconstructed CT scan $\hat{X}^n$ from the auto-encoder as input and feeds it into an encoder branch $E_t(\hat{X})$ that classifies the CT dataset/technology. In our experiments, $E_t$ relies on the VGG-Net architecture as well.

COVID-19 Classifier Module: The COVID-19 classifier $E_c$ is also uses several backbone architectures. Given a reconstructed CT scan $\hat{X}^n$, it outputs COVID vs. non-COVID predictions, i.e. $E_c(\hat{X}^n)$.

B. Loss Functions and Evaluation Metric

The complete loss function of PrepNet is based on the various terms presented in Figure 1. It comprises a reconstruction loss $L_{rec}$ and two classification losses $L_{pseu}$ and $L_{covid}:

$$L_{total} = L_{rec} + L_{pseu} + L_{covid}$$

Given the labeled dataset $D = \{(X^n_i, y^n_i, p^n_i)\}_1^N$ comprising the CT scans $X^n_i$ together with their binary COVID label $y^n_i$ and the CT-dataset pseudo label $p^n$, the auto-encoder reconstruction loss is given by $L_{rec} = \sum_i \|X^n_i - \hat{X}^n_i\|_2^2$; the COVID-19 binary classification loss is denoted $L_{covid} = -\sum_i y^n_i \log \hat{y}^n_i + (1 - y^n_i) \log (1 - \hat{y}^n_i)$; the CT dataset pseudo label is computed by $L_{pseu} = -\sum_i p^n_i \log \hat{p}^n_i$.

To measure the COVID-19 detection performance and to minimize the effect of class imbalance in datasets, we use the balanced accuracy metric (BA) [32]

$$BA = \frac{TP}{P} + \frac{TN}{N}$$

where $P$ and $N$ are the number of positive and negative samples respectively and $TP$ and $TN$ denote the number of true positive and true negative predictions, respectively. In addition, we also use specificity, sensitivity, and area under the curve to evaluate the COVID-19 performance results.

IV. EXPERIMENTS

A. Datasets

We use three public datasets to validate our approach experimentally. The SARS-CoV-2 CT-scan dataset [11] comprises a total of 4,173 CT images of real patients from the Public Hospital of the Government Employees of Sao Paulo (HSPM) and the Metropolitan Hospital of Lapa, both in Sao Paulo - Brazil (2,168 positive/infected and 768 healthy patients). Moreover, 1,247 CT scans belong to patients who have other pulmonary diseases. The CT image annotations (positive vs. negative) have been done by three different clinicians. Note that during our visual inspection we found two erroneous images (i.e. unrelated to the problem domain) and excluded them from the dataset. In addition, we also excluded the 1,247 pulmonary diseased patients.

The UCSD COVID-CT dataset [12] has been collected in the Tongji Hospital in Wuhan, China during the outbreak of COVID-19 between the months of January/2020 and April/2020. This dataset contains 349 CT images from infected patients and 397 from non-infected patients. All images have
be annotated by a senior radiologist of the same hospital. As reported by [22], heights of the images in this dataset range between 153 and 1,853 pixels with an average of 491 pixels, whereas the widths vary between 124 and 1,458 pixels (average of 383 pixels). For partitioning, we follow the splitting guideline provided by the authors of the dataset. Table I summarizes the train, validation and test splits for each dataset.

The MosMed dataset [13] was collected by the Moscow Health Care Department from different municipal hospitals in Russia between March/2020 and April/2020. The dataset contains axial CT images from 1110 patients with different levels of COVID-19 severity, ranging from mild to critical cases and also healthy patients. Some image samples of each dataset are provided in Figure 2.

The results in Table II show that the average cross-dataset performance (over all dataset splits) of models trained on original data increases by 0.77pp after using the pure auto-encoder model, and by 11.84pp through PrepNet. However, the average test accuracy for within-dataset evaluation declines by 0.32pp and 1.83pp after applying the baseline auto-encoder or PrepNet, respectively. A discussion regarding this effect is presented in the next section.

In our experiments, we use the VGG19 [16] as the baseline model because it is more straightforward to train and has shown good generalization properties on 2D medical images based on previous practical experiments1. Besides that, the VGG architecture has been also successfully applied for COVID-19 identification [2], [21].

As part of our ablation study, we also evaluated how different backbones affect the COVID-19 diagnosis accuracy of PrepNet. More precisely, we replicate the experiments for each backbone (SARS-COV-2 and UCSD COVID-CT) and evaluate different CNN architectures as part of our COVID-classifier Module (See Section III-A for more information). The CNN architectures include ResNet18 [17], Inception [35], and EfficientNet-B0 [36]. We report results in Table III. Experimental results show that in almost all backbones, the average cross-dataset performance increases with the cost of a small decrease in the within-dataset accuracy.

Finally, in order to evaluate the generalisation capabilities of PrepNet and our baselines, we evaluate how our trained models perform on an unseen dataset, i.e. the MosMed dataset [13]. The results in Table IV show the improvements of our AutoEncoder and PrepNet models in terms of BA and sensitivity, however, with a decrease in specificity and AUC when compared with the COVID-19 classifier. Despite the decrease in specificity, we argue that especially for medical diagnosis and screening, a low specificity is less harmful than a reduction in sensitivity, as false positive cases can be discarded by additional examinations. On the contrary, a higher sensitivity is important as false negatives should be low.

---

1https://stanfordmlgroup.github.io/competitions/mura/
TABLE I

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Size</th>
<th>Country</th>
<th>Train Percentage</th>
<th>Validation Percentage</th>
<th>Test Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARS-COV-2 [11]</td>
<td>2D CT</td>
<td>Various</td>
<td>Brazil</td>
<td>2,046 (70%)</td>
<td>439 (15%)</td>
<td>439 (15%)</td>
</tr>
<tr>
<td>UCSD COVID-CT [12]</td>
<td>2D CT</td>
<td>Various</td>
<td>China</td>
<td>425 (57%)</td>
<td>110 (16%)</td>
<td>201 (27%)</td>
</tr>
<tr>
<td>MosMed Dataset [13]</td>
<td>3D CT</td>
<td>Various</td>
<td>Russia</td>
<td>1100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE II
TEST PERFORMANCE OF DIFFERENT BASELINES COMPARED TO OUR PrepNet MODEL. RESULTS DEMONSTRATE THAT OUR MODEL IS CAPABLE OF INCREASING THE CROSS-DATASET AVERAGE.

<table>
<thead>
<tr>
<th>Dataset portion</th>
<th>BA</th>
<th>Sens</th>
<th>Spec</th>
<th>AUC</th>
<th>COVID classifier</th>
<th>Within Test Average</th>
<th>Cross-Dataset Average</th>
<th>Pre-trained encoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARS-COV-2</td>
<td>0.8924</td>
<td>0.9292</td>
<td>0.7876</td>
<td>0.8584</td>
<td>0.8587 (baseline)</td>
<td>0.4159</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>UCSD COVID-CT</td>
<td>0.3295</td>
<td>0.3476</td>
<td>0.2743</td>
<td>0.3110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SARS-COV-2</td>
<td>0.8056</td>
<td>0.9907</td>
<td>0.6460</td>
<td>0.8183</td>
<td>0.8555 (-0.32%)</td>
<td>0.4836 (+6.77%)</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>UCSD COVID-CT</td>
<td>0.49405</td>
<td>0.6630</td>
<td>0.3008</td>
<td>0.4519</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SARS-COV-2</td>
<td>0.9007</td>
<td>0.9353</td>
<td>0.7982</td>
<td>0.8668</td>
<td>0.9175 1.0067</td>
<td>0.5121 (-1.83%)</td>
<td>0.5343 (+11.84%)</td>
<td></td>
</tr>
<tr>
<td>UCSD COVID-CT</td>
<td>0.5545</td>
<td>0.6446</td>
<td>0.1858</td>
<td>0.4852</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

D. Discussion
The baseline and proposed pre-processing approaches introduce performance drops when applied before within-dataset classification. These approaches usually reduce the test accuracies when trained and evaluated on the same dataset using the corresponding dataset splits. Therefore, we further investigate the intermediate results of the baseline auto-encoder and PrepNet on a case-by-case basis. Severe cases of generated artifacts through reconstruction via the baseline auto-encoder and the PrepNet are presented in Figure 3. We conjecture that the drop in within-dataset test performance is caused by the intermediate results of the baseline auto-encoder and PrepNet. This information could be partially retained by propagating the gradients of the COVID-19 classifier network through the automatic encoder module. This way, it is possible to shift the focus of model training from merely optimizing hold-out test set performance on the same data distribution (which likely does not transfer to any other environment) towards cross-dataset detection accuracy. The proposed PrepNet improves the cross-dataset balanced accuracy by a margin of 11.84 percentage points (SARS-CoV-2 CT-scan dataset [11]) at the expense of a decline in the within-dataset test performance of ca. 1.83pp (UCSD COVID-CT database [12]). These results suggest that the trainable preprocessing network erases some of the necessary information for diagnosis, due to artifacts. This information could be partially retained by propagating the gradients of the COVID-19 classifier network through the preprocessing model, and generated artifacts could be detected automatically by monitoring the reconstruction loss of the auto-encoder module. This, together with further investigations on the applicability and generality of the proposed approach to combine multiple datasets, is an intriguing theme for future research.

V. Conclusions and Future Work
In this paper, we introduced a novel approach to unify several CT scan datasets with respect to varying image datasets and acquisition circumstances such as CT scanner technology through training an adaptive pre-processing network that removes such specificities from the images themselves. Additionally, we presented initial results demonstrating the applicability of the method on three publicly available benchmark datasets. This way, it is possible to shift the focus of model training from merely optimizing hold-out test set performance on the same data distribution (which likely does not transfer to any other environment) towards cross-dataset detection accuracy. The proposed PrepNet improves the cross-dataset balanced accuracy by a margin of 11.84 percentage points (SARS-CoV-2 CT-scan dataset [11]) at the expense of a decline in the within-dataset test performance of ca. 1.83pp (UCSD COVID-CT database [12]). These results suggest that the trainable preprocessing network erases some of the necessary information for diagnosis, due to artifacts. This information could be partially retained by propagating the gradients of the COVID-19 classifier network through the automatic encoder module. This, together with further investigations on the applicability and generality of the proposed approach to combine multiple datasets, is an intriguing theme for future research.

ACKNOWLEDGMENT
This research was financially supported by the ZHAW Digital Futures Fund under contracts “SDMCT—Standardized...”
### Table III
Experimental results of PrepNet with different backbones: VGG19 [16], ResNet18 [17], Inception [35], and EfficientNet-B0 [36]. Note that PrepNet increases the cross-dataset average.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Original</th>
<th>Baseline auto-encoder</th>
<th>PrepNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARS-COV-2</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
<tr>
<td>UCSD COVID-CT</td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
</tbody>
</table>

![Fig. 3. Severe cases of artifacts generated by the baseline and the proposed PrepNet. The images demonstrate different levels of distortions like e.g. extreme contrasts.](image7.png)

### Table IV

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pre-trained encoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARS-COV-2</td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td>UCSD COVID-CT</td>
<td><img src="image9.png" alt="Image" /></td>
</tr>
</tbody>
</table>

![Fig. 4. Samples CT scans that are wrongly classified after the trainable preprocessing.](image10.png)

---

Data and Modeling for AI-based CoVID-19 Diagnosis Support on CT Scans as well as “Synthetic data generation of CoVID-19 CT-X-rays images for enabling fast triage of healthy vs. unhealthy patients”.

---

**References**


Bias, awareness, and ignorance in deep-learning-based face recognition

Samuel Wehrli · Corinna Hertweck · Mohammadreza Amirian · Stefan Glüge · Thilo Stadelmann

Received: 30 July 2021 / Accepted: 4 October 2021
© The Author(s) 2021

Abstract
Face Recognition (FR) is increasingly influencing our lives: we use it to unlock our phones; police uses it to identify suspects. Two main concerns are associated with this increase in facial recognition: (1) the fact that these systems are typically less accurate for marginalized groups, which can be described as “bias”, and (2) the increased surveillance through these systems. Our paper is concerned with the first issue. Specifically, we explore an intuitive technique for reducing this bias, namely “blinding” models to sensitive features, such as gender or race, and show why this cannot be equated with reducing bias. Even when not designed for this task, facial recognition models can deduce sensitive features, such as gender or race, from pictures of faces—simply because they are trained to determine the “similarity” of pictures. This means that people with similar skin tones, similar hair length, etc. will be seen as similar by facial recognition models. When confronted with biased decision-making by humans, one approach taken in job application screening is to “blind” the human decision-makers to sensitive attributes such as gender and race by not showing pictures of the applicants. Based on a similar idea, one might think that if facial recognition models were less aware of these sensitive features, the difference in accuracy between groups would decrease. We evaluate this assumption—which has already penetrated into the scientific literature as a valid de-biasing method—by measuring how “aware” models are of sensitive features and correlating this with differences in accuracy. In particular, we blind pre-trained models to make them less aware of sensitive attributes. We find that awareness and accuracy do not positively correlate, i.e., that bias ≠ awareness. In fact, blinding barely affects accuracy in our experiments. The seemingly simple solution of decreasing bias in facial recognition rates by reducing awareness of sensitive features does thus not work in practice: trying to ignore sensitive attributes is not a viable concept for less biased FR.

Keywords Fairness · Convolutional neural networks · Discrimination · Ethnic bias · Gender bias

1 Introduction
FR has improved considerably and constantly over the last decade [1–4], giving rise to numerous applications ranging from services on mobile consumer devices to the use by law enforcement agencies [5–7]. The increased deployment has triggered an intense debate on the dangers of the pervasive use of biometrics [8–11] up to the point where regulation [12] and bans on the technology are discussed [13] and partially enforced [14, 15]. Several civil rights groups oppose facial recognition models as they can easily be used for mass surveillance [13].

Besides these fears of surveillance, another critical issue is that facial recognition tools have been shown to perform at different levels of accuracy depending on which socio-demographic group a subject belongs to. In a seminal study of commercial face recognition software, Buolamwini and

---

*Samuel Wehrli
samuel.wehrli@zhaw.ch

1 School of Social Work, Zurich University of Applied Sciences, Pfingstweidstrasse 96, 8005 Zurich, Switzerland
2 Institute of Data Analysis and Process Design, Zurich University of Applied Sciences, Winterthur, Switzerland
3 Centre for Artificial Intelligence, Zurich University of Applied Sciences, Winterthur, Switzerland
4 Institute for Applied Simulation, Zurich University of Applied Sciences, Wädenswil, Switzerland
5 Institute of Neural Information Processing, Ulm University, Ulm, Germany
6 Department of Informatics, University of Zurich, Zurich, Switzerland
7 ECLT European Centre for Living Technology, Venice, Italy

Published online: 27 October 2021
Gebru [16] showed that these tools tend to misclassify darker-skinned women more often than lighter-skinned men. As face recognition is increasingly relied on to grant individuals access to services and locations, and to predict people’s behavior, bias against certain socio-demographic groups easily results in these groups being more likely to be excluded from such services and locations. Bias becomes even more problematic when FR is used to identify suspects in a crime. When the FR algorithm misidentifies a person, this can have severe consequences: the misidentified person might unjustly be investigated or even charged with a crime they did not commit (as in the case of an American university student who was wrongly accused of terrorism by Sri Lankan police [17] or in the case of a black man who was wrongfully arrested in Michigan [18]).

Hence, the different levels of accuracy can be understood as an issue of bias: we expect FR to show approximately equal levels of accuracy for all socio-demographic groups and call it “biased” if it does not. One reason for the unequal levels of accuracy in FR is that the huge diversity in the appearance of human faces is not properly represent in the data used to train such models. Existing datasets tend to overrepresent lighter-skinned male faces, while other socio-demographic groups are underrepresented [16]. While bias also occurs when humans are the ones responsible for recognizing faces, the issue is more severe when conducted by machines as algorithmic decisions scale in speed, extent, and scope.

When one wants to avoid biased decisions made by humans, a standard approach is to make sensitive attributes unavailable. An every-day example are resumes: in the US, age and gender information as well as images are omitted in resumes. If the recruiters in charge are not aware of ethnicity, gender, and age—so the thinking goes—then decisions made by them cannot be biased by these sensitive features. Such biases could lead to strong candidates being wrongfully omitted. Ignorance of sensitive attributes is thus seen as a way of finding better candidates while mitigating discrimination. The methodology of being blind toward sensitive features is not new: As the “veil of ignorance,” it is part of John Rawls’s influential book “A Theory of Justice” (1971) which deals with the political philosophy of just distribution and fairness. Behind this “veil of ignorance,” people do not know their own identity and circumstances of life (gender, job, health, etc.). Rawls uses this concept as part of a thought experiment to find the principles based on which society and its institutions should be designed. Because people are biased by their situation in life (e.g., by knowing that they are born as a cisgendered white man), asking people for their ideas for such principles would most likely lead to biased principles. Therefore, Rawls asks people to imagine themselves behind this “veil of ignorance” in this newly constructed world. John Rawls demands ignorance of our own identity when imagining the “ideal” society and its institutions—with the goal of reducing bias and creating better results.

With both recruiting and the “veil of ignorance,” the assumption is that humans’ awareness of sensitive features leads them to make biased decisions, which harms marginalized groups. As Nyarko et al. [20] show, people are quick to apply this concept of removing sensitive attributes to machine learning models—despite the potentially harmful consequences for the disadvantaged group (see, e.g., [21–23]). The underlying assumption of those people is that removing sensitive attributes would reduce bias and thus help the disadvantaged group.

Considering the case of FR, this would mean that to reduce bias, “awareness” of sensitive features has to be removed. As stated above, we define bias as notably different level of accuracy between socio-demographic groups. We will refer to “awareness” as the extent to which a machine (e.g., a FR model) is aware of the presence of sensitive features (e.g., gender or ethnicity). In the literature, awareness is often equated with bias in FR: the idea is that removing awareness (i.e., decorrelating facial representations and sensitive attributes) simultaneously reduces bias—similar to how it is assumed that hiding sensitive features removes humans’ tendency to make prejudiced decisions [24–28].

In this paper, we explore what this removal of awareness means on a technical level and demonstrate why it cannot be equated with reducing bias. We thus argue that bias ≠ awareness with the important consequence that dealing with awareness in FR models does not necessarily reduce bias in any desired way. The rest of the paper is organized as follows: Sect. 2 describes how this work relates to other studies in this field. Section 3 explains the methods and data as well as the existing face recognition models that we use to experimentally examine the relationship of awareness and bias in face recognition. We then present and discuss the results of these experiments in Sects. 4 and 5 and draw conclusions in Sect. 6.

2 Related work

Racial biases are an issue across different sub-domains of computer vision: besides the field of FR, image classification models have, for example, been criticized for mis-labeling black men as “primates” [29]. Through the work of Joy Buolamwini and Timnit Gebru [16], FR’s bias
problem became a topic of public debate [30]. A follow-up study of their work revealed that Microsoft, IBM, and Face++ released new versions of their API that improved the audited metrics [31]. IBM also removed the facial detection from its API in September 2019. In the public sector, San Francisco was the first US city to ban the usage of FR technology in 2019 [32] as a consequence of discovering biases in FR [33]. Several other cities followed. Since then, a tremendous amount of research has measured [25] and reduced [34] biases in FR technologies. The remainder of this section presents recent works on measuring racial biases and methods for removing biases.

To reduce the racial bias, researchers followed these main directions: 1) balancing datasets, 2) model selection and loss design, and 3) removing the sensitive feature in the representations of faces used for identification. This paper relates to the third category and investigates the effect of a blinding method to remove sensitive features from face matching techniques. Systematic reviews of the recent attempts to tackle biases in machine learning and FR are represented in [35, 36].

Identifying and quantifying the amount of bias in FR technology are the initial step toward less-biased FR. Garcia et al. showed that face matching confidence of FR models correlates with gender and ethnicity, thus revealing demographic bias [37]. Cavazos et al. demonstrated that different thresholds are needed to equalize false accept rates (FARs) and the recognition accuracy [38]. Serna et al. quantified FR bias using normalized overall activation of the models for various races [39].

The first direction to overcome racial bias in FR is addressing the bias in the data: e.g., measurement error (systematic errors in the measurements of variables for specific groups) or representation bias (not everyone has the same probability of being in the dataset, meaning that the training data do not represent the real world’s diversity) [35]. To suppress the effects of imbalance in datasets, Kortylewski et al. proposed using synthetic data [40]. Robinson et al. introduce a racially balanced dataset [34]. With that, they were able to show how the performance gaps in FR for various races decrease when adapting the decision thresholds for each race.

Modifying the model choices, e.g., the training process and the target, is the second venue researchers explored to remove racial bias in FR [41]. Yu et al. adapted the selection of face samples for training based on the data distribution and model bias [42], while Wang et al. attempted to transfer knowledge from the source domain (Caucasian) to target domains (other races) through learning facial features with adequate generalizing across different races [43]. The last approach is removing sensitive information related to races [24]. Generative Adversarial Networks (GANs) inspired several researchers to blind models and/or reduce the correlation between sensitive features and facial attributes for face recognition [26, 27, 44, 45]. Adeli et al. proposed an adversarial loss to minimize the correlation between model representations and sensitive information (races) and statistical dependency of the learned features and source of bias (racial group) [46]. In this paper, we use a blinding technique that applies to every trained model to remove sensitive information from model representations and demonstrate that removing awareness does not necessarily remove the racial bias in several FR technologies.

3 Materials and methods

We start this section with a short recapitulation of the fundamentals of deep FR models. Afterwards we introduce the FR models and the evaluation dataset we chose for the present work. Finally, we detail the methods we used to quantify/measure the awareness of deep FR models regarding specific socio-demographic groups and to remove information in the models’ embeddings with respect to these groups.

3.1 Basics of deep face recognition models

Wang et al. [47] provide a comprehensive survey on deep FR methods, including algorithms, databases, training protocols, and applications. In this section, we limit ourselves to a short review of the basic algorithm that yields a compact representation of a face in a feature space, a.k.a. embedding.

In image processing, deep learning methods, such as Convolutional Neural Network (CNN), use a cascade of multiple layers of processing units for feature extraction and transformation. It was shown that each layer learns multiple levels of representations which correspond to different levels of abstraction depending on the task at hand. Regarding face recognition, a major advantage of this hierarchy of concepts is a strong invariance to changes in face pose, lighting, and expression changes. Figure 1 shows the general structure of a CNN on the task of face identification, together with the features learned on different levels of the network hierarchy. Contrastive and triplet loss functions are used to train the CNNs of deep FR models [2]. They optimize the CNN, such that embeddings (i.e., features) of positive image pairs (same identity) are close to each other, whereas embeddings of negative pairs (different identity) are pushed apart.

To summarize, every face image \( I (160 \times 160 \text{pixel corresponding to 25,600 features}) \) is processed in a deep CNN that

learned a hierarchy of features ranging from basic concepts like edges up to complex concepts like eyes. On top of these features, a compact vector representation $x$ (typical length of 128) is learned. In general, we refer to these compact representation as embedding. Furthermore, the model is trained to generate embeddings, such that $x_i$, $x_j$ are close together.
if the input images \( I_i, I_j \) belong to the same person. Figure 2 depicts this concept.7

### 3.2 Face recognition models

As described above, we are operating on the embedding space of FR models. This allows us to easily compare different readily-trained models. For our comparison, we chose two different openly available FR models/architectures: first, a model from the popular Visual Geometry Group (VGG) [48] family and second, FaceNet [2].

Together with the VGGFace2 dataset [48], the authors provide trained models8 that we applied in our study. The models were pre-trained on the MS-Celeb-1M [49] dataset and then fine-tuned on VGGFace2. These SE-ResNet-50 models follow the architectural configuration in [50]. Besides the architecture, the models differ in the lower dimensional embedding layer (128D/256D) which is stacked on top of the original final feature layer (2048D) adjacent to the classifier. All models were trained with standard softmax loss. In our experiments, we did not see any significant difference regarding different sizes of the embedding layer. Hence, we show the results for the VGG128 with an embedding layer of size 128.

FaceNet is a face recognition system that was described by Florian Schroff, et al. at Google [2]. In our experiments, we use a model that is based on the GoogLeNet style Inception models [51] with approximately 23M trainable parameters. The trained model is freely available4 and was pre-trained on the MS-Celeb-1M [49] dataset. The embedding layer is also of size 128.

### 3.3 Evaluation dataset to study racial bias

In the following, we look at the sensitive attributes ethnicity and gender. The Racial Faces in-the-Wild (RFW) dataset was designed to study racial bias in FR systems [43]. Images are annotated with one of four labels, namely Caucasian, Asian, Indian, and African, which form four ethnic clusters. Each subset contains about 10k images of 3k individuals for face verification. According to [43], the labels for Caucasian and African ethnicity were assigned by the Face++ API [52] and those for Asians and Indians from the nationality attribute in FreeBase celebrities [53]. To avoid the negative effects caused by the biased Face++ tool, the authors manually checked some images with low confidence scores from Face++.

### 3.4 Awareness of sensitive features

We investigate how well machine learning models can predict the sensitive features, such as ethnicity and gender, based on the face embedding. The intuition is that an FR model is “aware” of a sensitive feature if it can be predicted from the embedding vectors produced by the FR model. This inference is a classification task and the performance depends on the classification model at hand. If simple models, more precisely models with a low number of parameters, can properly infer the sensitive features, awareness is high. If models with a high number of parameters are required, then awareness is lower. There is no awareness if the features cannot be better predicted than random guessing.

### 3.5 Blinding

As stated above, deep FR models map images of faces into an embedding vectors \( \mathbf{x} \). Given a data set with labels for sensitive attributes such as ethnicity and gender, the images can be grouped into clusters based on these labels. To investigate the influence of this clustering on the face recognition rates, we propose a blinding procedure to remove the information related to the separation of these clusters in the

---

5 [Attribution: https://www.pinterest.ch/pin/663999538792168278/, CC BY-SA 4.0, via Wikimedia Commons.](https://www.pinterest.ch/pin/663999538792168278/)
6 [https://github.com/ox-vgg/vgg_face2.](https://github.com/ox-vgg/vgg_face2)

---

Table 1 Number of samples per cluster regarding different facial characteristics in the RFW dataset that are associated with bias

<table>
<thead>
<tr>
<th></th>
<th>Caucasian</th>
<th>Indian</th>
<th>Asian</th>
<th>African</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>6921</td>
<td>7419</td>
<td>5647</td>
<td>10,053</td>
<td>30,040</td>
</tr>
<tr>
<td>Female</td>
<td>3178</td>
<td>2802</td>
<td>3955</td>
<td>344</td>
<td>10,279</td>
</tr>
<tr>
<td>Total</td>
<td>10,099</td>
<td>10,221</td>
<td>9,602</td>
<td>10,397</td>
<td>40,319</td>
</tr>
</tbody>
</table>

The dataset is balanced with respect to ethnicity but skewed with respect to gender (75% male, 25% female). African women are strongly underrepresented and constitute less than 1% of all samples.

In addition to ethnicity, we added a gender label to each image using a Wide Residual Network trained on the UTK-Face [54] and IMDB-WIKI [55] datasets.5 The model’s performance is reported to be around 88% accuracy [56, 57]. The gender prediction is accessed in terms of a continuous score \( s_{\text{gender}} \) between 0 and 1, where lower values indicate male and higher values indicate female. To create fixed clusters of samples, the gender score was split at \( s_{\text{gender}} < 0.5 \) for male and \( s_{\text{gender}} > 0.5 \) for female. As we have multiple face images for each person in the dataset, each person was labeled to be male/female based on their mean score. Table 1 gives an overview of the resulting number of samples per cluster with respect to ethnicity and gender.
embedding space. The procedure is a linear operation and uses the following steps:

1. Compute the centroids of the clusters defined by the sensitive attributes in the embedding space.
2. Use a one-vs-rest (OvR) approach to calculate the “directions of discrimination” given by the centroids of each cluster relative to the other centroids.
3. Apply singular value decomposition (SVD) on the “directions of discrimination” to find an orthonormal basis spanning the “discriminatory subspace.”
4. Remove projections onto the “discriminatory subspace” from the embedding vectors. This results in new embedding vectors whose centroids fall on top of each other, i.e., the separation due to sensitive attributes is removed—hence the term “blinding”.

The procedure outlined above operates on the embedding vectors \( \mathbf{x}_i \), where \( i \) denotes the sample. Associated with each sample is a cluster label \( k \in \{1, \ldots, K\} \). As a first step, we define the centroids of each cluster by the average

\[
\bar{x}_k = \frac{1}{n_k} \sum_{i \in C_k} x_i, \tag{1}
\]

where \( C_k \) is the set of embedding vectors associated with cluster \( k \) and \( n_k \) is the corresponding size. Following an OvR approach, the normalized direction of discrimination of each cluster \( k \) to the other clusters is given by the vectors

\[
u_k = \frac{v_k}{\|v_k\|} \quad \text{with} \quad v_k = \bar{x}_k - \frac{1}{K-1} \sum_{i \neq k} \bar{x}_i, \tag{2}
\]

where \( K \) is the number of clusters. By construction, the vectors \( u_k \) are not linearly independent, but span a subspace of rank \( K-1 \). This can by verified by applying a SVD on the matrix \( U = [u_1 \ldots u_K] \). SVD also provides an orthonormal basis \( B = [e_1 \ldots e_{K-1}] \) of the corresponding subspace. The final step is to remove the projections onto this subspace by

\[
x_b^k = x_i - \sum_{j=1}^{K-1} (x_i \cdot e_j) e_j, \tag{3}
\]

where \( (x_i \cdot e_j) \) is the dot (or scalar) product. Equation (3) yields new embedding vectors \( x_b^k \) with the same shape as the original ones. The upper index \( b \) stands for “blinded” inspired by the fact that information with regard to the discriminatory dimension has been removed.

### 3.6 Experimental setup

We extract the embeddings of the approximately 40k face images from the RFW testset for the VGG and FaceNet models. Face detection and alignment is done using the MTCNN approach proposed by Zhang et al. [58]. Based on the embeddings, we analyze the models’ awareness regarding sensitive attributes ethnicity and gender using the classifier performance within the embedding space (cf. Sect. 3.4). Furthermore, we report the bias of the models based on the actual FR rates. In a second step, we apply the proposed blinding procedure (cf. Sect. 3.5) to the embedding spaces and again report the models’ awareness and bias.

### 4 Results

In this section, we present the results of our experiments. Specifically, we compare the model’s awareness of sensitive features as well as bias with respect to ethnicity and gender before and after the blinding procedures.

#### 4.1 Awareness of sensitive attributes

To visualize the structure of the embedding space (128 dimensions), we use t-distributed Stochastic Neighbor Embedding (t-SNE) [59] as an unsupervised way to reduce the dimensionality to a 2D representation. t-SNE maps points from a high-dimensional space to a lower dimensional space, preserving local distances. Points which are close to each other in the high-dimensional space remain close to each other in the low-dimensional space. This property is particularly suitable for the FR task where images are mapped to the embedding space and decisions are based on distances in this space. Figure 3 shows the 2D visualizations of the embeddings of the RFW test set. As one can see, the dimensionality reduction reveals well-separated clusters defined by ethnicity and gender. The fact that the data are skewed with respect to gender is reflected in the figures, in agreement with the sample counts shown in Table 1. The t-SNE visualizations for the VGG128 and FaceNet models are surprisingly similar: Caucasian men lie in the center, surrounded by the remaining groups in similar positions. As stated, t-SNE is an unsupervised clustering method and groups embedding vectors solely based on their respective distances without any direct information about the sensitive features. Nonetheless, the clusters appear very nicely separated, suggesting that the embedding space is separated into different sectors corresponding to different ethnicities and genders and that the models under considerations are therefore highly “aware” of the sensitive features.

To associate this vague intuition of “awareness” with a more formal approach, we investigate how well different classifiers predict the sensitive features based on the face embeddings. The intuition is that “awareness” toward

---

6 https://github.com/YYuanAnyVision/mxnet_mtcnn_face_detection.
sensitive features is high if simple models (low number of parameters) can accurately predict these features. The ability to predict sensitive features based on the embedding vector is evaluated for different classifiers. In the case of ethnicity, it is a multi-class classification task (4 classes). Therefore, we use macro-averaged F1-scores as a performance measure, which is the unweighted mean of the F1-scores of each label, as shown in Fig. 4. The upper bound of this score is 1. A lower baseline is given by random guessing and is the inverse of the number of clusters for a given sensitive feature, i.e., 0.25 for ethnicity and 0.5 for gender. Figure 4 shows that the classification scores on the original embeddings are close to 1. Moreover, simple linear classifiers such as nearest centroid classification and logistic regression (both with a low number of parameters) show the same performance as the more advanced neural network classifiers. This confirms the hypothesis made above that the embedding space is structured into sectors given by the sensitive features. The fact that centroid classifiers work well means that these sectors are linearly separable and well-represented by their centroids. Therefore, we conclude that the models are indeed “highly aware” of ethnicity and gender.

Previous work [25] investigated this structure by means of various clustering scores such as the Silhouette coefficient.

Fig. 3 t-SNE visualization (2D) of the embedding space of the RFW test set samples. The coloring is based on different labels corresponding to the sensitive features ethnicity (color) and gender (light versus dark). Left: VGG128 model. Right: FaceNet model (color figure online).

Fig. 4 “Awareness” as represented by the macro-averaged F1-scores for various classifiers, predicting sensitive features based on the face embeddings for the VGG128 model (upper panel) and the FaceNet model (lower panel). The following classifiers from scikit-learn were used: (Random) theoretical value of random guessing; (Centroid) nearest centroid classifier; (Logit) logistic regression; (NN-1,2) neural network with one or two hidden layers (100 nodes) and relu activation. A train/test split of two-thirds/one-third was used. The upper rows represent the scores using the original embeddings. The lower rows show the scores for the blinded embeddings. The scores are colored, ranging from red for random guessing to green for a perfect score of 1 (color figure online).
These scores suggested negligible structuring in contrast to present findings. The discrepancy is due to the high dimensionality of the embedding space: In order for the linear classifiers to work, it is enough if the sectors related to ethnicity and gender are separated by a few dimensions out of many (128 in the models under considerations). In contrast, clustering scores give an average clustering for all dimensions and yield low scores if only a few dimensions contributed to the separation, which explains the discrepancy.

The fact that the cluster centroids are representative for the whole cluster lends itself for further analysis. The positions of the centroids in the high-dimensional space can obviously not be visualized. It is however interesting to look at the radii (norm) from the origin of the embedding space to the centroids. This is depicted on the left side of Fig. 5 for the ethnic clusters. The embedding vectors for different faces all have a radius 1, because they are the output of a soft-max layer. The centroids, which are averages of embedding vectors, therefore have radii < 1 as can be seen in the figure. Interestingly, the radius of the Caucasian cluster is roughly half of the other radii, confirming that Caucasians are indeed closer to the origin which is inline with the t-SNE visualization shown in Fig. 3. The centroid classifier can also be used to generate a measure of the relative “size” of the different sectors by randomly generating embedding vectors (with radius 1) and classifying them with the centroid classifier. The result is shown on the right side of Fig. 5 for ethnicity. It turns out that the Caucasian sector covers roughly 90% of the embedding space in this metric. This is a clear and remarkable result which shows that different ethnic groups are treated differently in the models under consideration. To put it dramatically: In the models under consideration, 90% of the embedding space serves the purpose of Caucasian face recognition.

The differences in centroid radii and the embedding space coverage may be related to the bias in face recognition discussed below. The relation can either be causal (the difference in radii causes the bias) or rather indicative (the difference is due to the same root cause as the bias). To differentiate between the scenarios, we propose the blinding procedure described above which removes the dimensions separating the centroids (cf. Sect. 3.5). In the case of ethnicity, there are 4 centroids which span a 3D subspace (similar to the well-known fact that three points span a 2D plane). By blinding, i.e., projecting out this 3D subspace, the centroids fall on top of each other. As a consequence, the sectors corresponding to different ethnicities will be shifted together and the clusters are no longer represented by their centroids. We apply this blinding procedure to both ethnicity and gender. Figure 4 shows that the performance of the linear classifiers drops to random guessing as expected. Hence, the clusters are no longer linearly separable. The more complex neural network classifiers are still able to predict the sensitive features. Therefore, after blinding, more complex models with a higher number of parameters are needed to predict the sensitive features. This means, in our terminology, that the “blinded” embeddings are indeed less “aware” of the sensitive features.

4.2 Bias with respect to sensitive attributes

Face recognition rates and bias are evaluated with the RFW dataset. The RFW dataset provides image (i.e., embedding) pairs. These pairs can either be two pictures of the same person (positive pair) or of two different people (negative pair). The resulting task is a binary classification of the pairs into positive and negative pairs. Note that all pairs have the same ethnicity label (only people with the same ethnicity label are compared). The recognition rate is the accuracy of the corresponding classification. Here, we investigate the error rate

$$\text{Error rate} = 1 - \text{recognition rate} = \text{share of false positives} + \text{share of false negatives},$$

where
which has two contributions due to the two types of misclassification:

- false negatives: same identity predicted as different ones (red in the figures);
- false positives: different identities predicted as the same one (blue in the figures).

The classification itself is done by calculating the (Euclidean) intra-pair distance $d$ between the embedding vectors of the images of the pair and comparing it to a threshold $d_c$. A positive (same) pair is predicted for $d < d_c$. Otherwise, a negative (different) pair is predicted. The error rates of these classifications are shown in Fig. 6 for both the original and the blinded embeddings. A substantial difference is apparent between the different groups and to a marginal extent for the gender groups. The error rate is lowest for the label “Caucasian” ($\approx 10\%$), whereas the error rate is highest for the label “African” ($\approx 15\%$). Moreover, for people with the label “Caucasian”, false negatives are the most common error type. For people with the label “African”, however, false positives are the most common error type. Apparently, we observe two types of bias:

1. difference in the total error (or recognition) rate;
2. difference in ratio between false-positive and false-negative error rates.

As can be seen, removal of the separation of the sectors associated with different sensitive features (i.e., blinding) affects the error rates and bias only marginally. We conclude that the concept of awareness introduced above is different from bias. In addition, the difference in the centroid radii mentioned above does not cause the bias, as removing it by blinding does not affect it. t-SNE allows us to visualize where the misclassifications are located in the embedding space. Figure 7 shows that the misclassifications are randomly scattered. It is therefore not surprising that moving the different clusters on top of each other by blinding has only marginal effects on the recognition rate and bias.

### 5 Discussion

The results show that putting FR behind a “veil of ignorance”—where predicting people’s assigned gender and ethnicity label becomes harder—does not have any notable influence on the accuracy of the FR model. In this way, dealing with biases in machines is strikingly different from dealing with human biases in, e.g., job application screening where ignorance is often deemed to be a necessary condition for unbiased decision-making. We argue that this is because the task of screening job applicants is fundamentally different from the task of FR technology, which is to decide whether two pictures show the same person. If humans take biased decisions when looking at CVs (e.g., accepting more
men than women without any reasonable justification), we assume that this is caused by implicit or explicit prejudices. When humans have to match faces, biases in the form of lower performance in recognizing members of certain groups might not be caused by prejudices, but by a lack of exposure to these faces—the so-called “cross-race effect” [60]. The same likely applies to the FR models in our experiments: the reason for their different performance levels is not awareness of groups’ race or gender. Rather, it is because they have not “seen” as many people from that group before. We thus caution machine learning experts against trying to solve the issue of bias in FR by applying their intuition of fairness to FR. Simply ignoring sensitive attributes does not seem to be a proper mediator of unbiased FR. Instead, FR developers noticing biases in the performance rates of their models should avoid trying to “blind” their model, but should rather improve the training data of their model as suggested in previous work [16, 43].

However, the experiments also demonstrated how easy it is to predict the assigned gender and ethnicity label from the existing models, such as FaceNet. This is particularly worrisome considering recent reports of Russia deploying tools that detect people’s ethnicity [61] and China using tools that detect Uighur faces [62]. One might therefore want to consider if such open-source models should be “blinded” before their release. While it does not improve the accuracy rates for any groups, it also does not harm the accuracy, but might potentially protect from such tools being used to assign ethnicity labels to people.

We also found that there is an imbalance in the type of errors that the models make. We assume that the aforementioned imbalance in the training data is what leads to this difference in the types of errors that the model makes for the different ethnic groups. Assuming that the training data consisted of a disproportionately high number of Caucasian faces, the trained model would essentially provide more “space” in the embedding space for Caucasian faces. As noted in Sect. 4.1, the Caucasian sector indeed covers a large part of the embedding space, namely about 90%. This means that two Caucasian faces will on average have a greater distance to each other than two pictures from another group. When we now use the same threshold to determine at which distance two pictures are classified as “identical” instead of “different”, Caucasian faces—being generally more spread out—are more likely to be classified as “different”. This explains the high false-negative rate for the Caucasian group. Pictures from other groups are generally closer to each other, which makes a classification of two pictures as showing the same person more likely and explains the comparatively higher false-positive rate. As suggested by Robinson et al. [34], a way to deal with these different levels of error types is to create ethnicity-specific thresholds for when two faces are classified as “identical”. This distance up to which Caucasian faces are classified as “identical” would have to be higher than that of other groups. However, this requires a

---

**Fig. 7** This figure is derived from the t-SNE coordinates shown in Fig. 3. Each point in the plot represents the average coordinates of a pair, either positive with the same identity or negative with different identity. The grey points represent correct predictions of positive (same identity) or negative (different identity) pairs by the face recognition algorithm. Red points are the cases where positive pairs are mistakenly classified as negative pair. Blue points are the cases where negative pairs are mistakenly classified as positive pair. Left: VGG128 model. Right: FaceNet model (color figure online)
classification of individuals into ethnic groups before checking potential matches, which is a at least questionable practice considering how the results of such an analysis could be misused. Therefore, while this solution makes sense on a technical level, we question if its benefits (equalization of error types across groups) outweigh the potentially harmful consequences (misuse of ethnicity classifications). Insofar, it once again appears to be the better option to address the root cause of the issue of the differing error rates: the training set.

The reason why the differences in error types can have dire consequences becomes evident when we think about a common application of facial recognition: policing. False negatives in the context of policing might mean not recognizing a suspect as the camera footage of them is not matched with their picture in the database. False positives are cases where camera footage of an innocent person is confused with pictures of the wanted person. As false negatives are far more common for Caucasian faces than false positives, the FR system errs on the side of not flagging Caucasian people as potential suspects. The situation is reversed for faces labeled as “African”: here, the system is more likely to flag an innocent person as a suspect than it is to accidentally let a suspect pass by without being flagged. This is particularly problematic considering the discrimination that people of color face in many regions and, on the other hand, the privilege that white people have. Anna Lauren Hoffmann points out “the different real-world consequences false results might have for different groups” [63, p. 907] in the context of AI applications. “A person of relative socioeconomic advantage is more likely to have the time or resources necessary to contest an unfair decision—an imbalance that persists regardless of the fact that differently-situated groups stood an equal chance of being falsely flagged within the system” [63, p. 907]. When we are looking at error types for marginalized groups, we thus also have to consider how these groups deal with the different error types.

Finally, we want to emphasize that this paper only considered one of the two main issues of FR technology: its biases, i.e., different levels of accuracy and different error types. While we showed that a seemingly easy solution to this problem (“blinding” the model) does not work in practice, even a method that mitigates this issue would not address the second and arguably more pressing issue of FR: its potential usage for mass surveillance. Simply advocating for better performance rates might be dangerous as long as there are no policies that guide what FR can legally be used for. Moreover, we have to keep in mind that mass surveillance, just like other harmful technologies, disproportionately affects marginalized groups. Mohamed et al. [64], for example, describe several cases in which problematic AI applications are beta-tested in poorer communities as less resistance is expected due to their limited resources. An example of this is the case of a facial recognition systems that has been proposed to be used in a Brooklyn apartment complex whose tenants are mostly black. The tenants worry that the technology is not implemented for a safer environment for the tenants, but rather for their surveillance as such data could be abused [65]. Considering the potentially harmful consequences of FR technology as well as the systemic disadvantages marginalized communities face, simply advocating for equal accuracy rates is thus not enough. Instead, we need to think further about how such technologies are being used. This includes policy as the previously mentioned bans on facial recognition (see, e.g., [32]) and the EU’s current attempt of regulating AI [66, 67].

6 Conclusion

In this paper, we operationalized “awareness” as a measure of how well different classifiers predict the sensitive features based on the face embeddings. The intuition is that “awareness” toward sensitive features is high if simple models (with a low number of parameters) are sufficient for accurately predicting the sensitive features based on the face representations learned by an FR neural network. For example, we would say that a model’s “awareness” of gender is high if a simple linear model can accurately predict the gender of a test subject from its picture when trained on face embeddings and gender labels.

Inspired by the example of human job application screening and its early adoption in the FR literature, we introduced a blinding procedure to reduce awareness. We showed that this procedure allows us to reduce awareness in a controlled and selective way. Applying this procedure enabled us to answer the question of whether removing information about sensitive features helps to reduce bias as it is assumed to do in the human example. For this, we compared the models’ awareness and bias before and after blinding. We came to the conclusion that indeed bias ≠ awareness.

We further found that the models make different kinds of errors for different ethnic groups. For the Caucasian group, the models are more likely to identify two pictures of the same person as different people. The opposite is the case for faces labeled as “African.” Again, blinding the models did little to change that. Instead, improving the training data seems to be the method that is more reliable when it comes to reducing bias in FR.

6.1 Limitations

Our study comes with certain limitations. The main limitation is the data used for our experiments. We tried to use a well-balanced dataset in terms of ethnic groups and gender groups. We picked the RFW dataset as it consists of approximately equally sized ethnic groups. However,
the dataset does not contain labels for other attributes such as “gender.” We therefore had to annotate the dataset ourselves, for which we used a pre-trained classifier. We found that faces labeled as “female,” and in particular such that are also labeled as “African,” are underrepresented in the RFW dataset. It is thus unclear if our results are representative of these underrepresented groups (e.g., faces labeled as “female” and “African”).

Regarding the calculated error rates, we only compared pairs of faces with equal labels, e.g., “female” to “female” and “Indian” to “Indian”. We have not tested how likely the models are to make false-positive errors when given two faces with different labels. However, the false-positive rates of confusing faces from group A with faces from group B would be the same as the other way around. Therefore, such an analysis would be unlikely to give us new insights. The interesting analysis is thus mainly the in-group comparison that we focused on.

### 6.2 Future work

Importantly, considering the potentially harmful consequences of FR, future work should ask in what situations FR technology is inappropriate to use and the question of how misuse of these systems can be prevented, e.g., through policies. Only once these questions have been addressed can advocacy for less-biased FR be beneficial to marginalized communities.

As discussed, the easiest and least problematic way to improve error rates seems to be to improve the data that are used to train the FR models. This means ensuring that sensitive attributes and their intersections are more equally represented. Respective research would then need to confirm that this is actually sufficient. Existing work by, e.g., Buolamwini and Gebru [16], already created a dataset that is diverse in terms of skin color and gender. Future work could continue on this path and ensure diversity along other axes.

### Code availability

The source code is available and accessible via the following link: [https://github.com/samuelwehrli/Face-Recognition-Bias](https://github.com/samuelwehrli/Face-Recognition-Bias).

### Declarations

**Conflict of interest** The authors have no conflict of interest concerning this research.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).

### References

41. Hooker, S.: Moving beyond algorithmic bias is a data problem. Patterns 2(4), 100241 (2021)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Customizing skills for assistive robotic manipulators, an inverse reinforcement learning approach with error-related potentials

Iason Batzianoulis1,5, Fumiaki Iwane1,2,3,5, Shupeng Wei1,5, Carolina Gaspar Pinto Ramos Correia1, Ricardo Chavarriaga2, José del R. Millán2,3,4,6,7 & Aude Billard1,6

Robotic assistance via motorized robotic arm manipulators can be of valuable assistance to individuals with upper-limb motor disabilities. Brain-computer interfaces (BCI) offer an intuitive means to control such assistive robotic manipulators. However, BCI performance may vary due to the non-stationary nature of the electroencephalogram (EEG) signals. It, hence, cannot be used safely for controlling tasks where errors may be detrimental to the user. Avoiding obstacles is one such task. As there exist many techniques to avoid obstacles in robotics, we propose to give the control to the robot to avoid obstacles and to leave to the user the choice of the robot behavior to do so a matter of personal preference as some users may be more daring while others more careful. We enable the users to train the robot controller to adapt its way to approach obstacles relying on BCI that detects error-related potentials (ErrP), indicative of the user’s error expectation of the robot’s current strategy to meet their preferences. Gaussian process-based inverse reinforcement learning, in combination with the ErrP-BCI, infers the user’s preference and updates the obstacle avoidance controller so as to generate personalized robot trajectories. We validate the approach in experiments with thirteen able-bodied subjects using a robotic arm that picks up, places and avoids real-life objects. Results show that the algorithm can learn user’s preference and adapt the robot behavior rapidly using less than five demonstrations not necessarily optimal.

https://doi.org/10.1038/s42003-021-02891-8

1 Learning Algorithms and Systems Laboratory (LASA), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. 2 Brain-Machine Interface (CNBI), École Polytechnique Fédérale de Lausanne (EPFL), Geneva, Switzerland. 3 Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, USA. 4 Department of Neurology, University of Texas at Austin, Austin, TX, USA. 5 These authors contributed equally: Iason Batzianoulis, Fumiaki Iwane, Shupeng Wei. 6 These authors jointly supervised this work: José del R. Millán, Aude Billard. 7 Email: jose.millan@austin.utexas.edu, aude.billard@epfl.ch
Individuals with spinal cord injury (SCI) often experience permanent neurological deficits and severe motor disabilities, which impair their ability to perform even the simplest everyday tasks, such as reaching-to-grasp objects. Assistance from robots may enable patients to recover some of their lost dexterity by letting a robotic system to perform these tasks on their behalf. In those cases where residual muscular capabilities are not reliable enough, a way to control such assistive robotic system is through brain–computer interfaces (BCIs)1,2.

A BCI measures and decodes the subject’s neural activity, translating their motor intention into the corresponding actions of a robot arm. These BCIs decode cortical correlates of movement parameters such as velocity3–5 or position6,7, thus providing direct control of the robot arm. Nevertheless, even after rather long training sessions, BCI performance still suffers from large variability over time and is significantly slower than the corresponding human actions.

Moreover, operation of the system relies on a continuous modulation of the brain signals to control the robot’s motion. Such intense level of concentration may not be amenable to all users. Involuntary changes in the user’s mental state, as well as fatigue and workload, may deteriorate BCI performance8,9.

Such continuous modulation of the brain signals is therefore meant to be imprecise and cannot be used reliably for tasks that require fast reactivity and high precision, such as when avoiding obstacles. Hence, to facilitate user’s learning and control, we propose to grant some authority to the robotic system, by developing a shared-control paradigm for obstacle avoidance10, which exhibits high reactivity even in the presence of fatigue and workload, may deteriorate BCI performance8,9.

Motion planning has reached a high level of maturity for the control of robot arms11,12. A branch of motion planning is trajectory generation driven by dynamical systems (DS), where robot’s reaching motion toward a target is modeled through a vector field with one attractor located at the target (Fig. 1a). Once the target is defined, the robot’s velocity depends only on its position with respect to the target. The benefits of this method are, among others, that it enables real-time adaptation of the robot’s trajectory11. This makes it a natural framework for obstacle avoidance12, which exhibits high reactivity even in the presence of moving obstacles13.

Such a framework would avoid BCI depending solely on the user’s brain signals to drive the robotic arm, as the robot trajectories could be generated autonomously. However, these automatic trajectories might not be acceptable from the user’s perspective: the robot may approach the obstacle too closely or avoid it too sharply for their liking. We, therefore, propose a method that combines robot learning and BCI techniques by which the user can train the system to learn the obstacle-avoidance behaviors that suit her/his individual preferences.

A critical component of our approach is how to gather users’ preferences. Consider a situation where the default trajectories bring the robot very closely to the object (Fig. 1b). From where the user stands, this motion may appear as too risky or even leading to a collision. The user does not need to indicate such a perceived misbehavior explicitly, something that people suffering from severe motor disabilities can hardly do; it can be detected directly from the user’s error-related potentials (ErrP)14,15. Error-related potentials are time-locked brain potentials elicited when actions do not match users’ expectations16–18. ErrPs are employed in BCIs for correcting or adapting brain decoders19,20 and have recently been introduced for direct control21–23. When in place, ErrPs were triggered by robot actions that are erroneous according to some explicit criteria, here we show that ErrPs are also elicited by error expectation—i.e., robot actions during its continuous movement that the user considers will lead to erroneous trajectories because they will not meet the user’s preferred obstacle avoidance behavior.

In our framework, upon detecting the occurrence of ErrPs, the system adjusts its control policy to generate future trajectories that may better fit the user’s implicit reward function. For this purpose, we rely on inverse reinforcement learning (IRL), an approach that uses demonstrations from experts to both learn a reward function and to produce the optimal trajectory according to this reward25–28. IRL can hence be used in conjunction with ErrPs to determine when and how to update the intelligent robot controller, as illustrated in Fig. 2.

We validate this ErrP-IRL approach with 13 subjects in two series of experiments, as illustrated in Supplementary Fig. 1. In a first experiment, 8 subjects interacted with the robot arm via minimal commands delivered with a joystick to start the robot’s motion (deflecting the joystick toward the rough desired direction) and to signal error expectation (releasing the joystick). After the onset of the motion, the robot moved autonomously from left to right, or vice versa, using its dynamical system to avoid a wine glass sitting in the middle of the trajectory. We show that our IRL-ErrP approach derives the preferred trajectories for each subject. Then, in a second experiment, 5 additional participants used the same joystick to make the robot arm perform pick-and-place tasks, while avoiding obstacles, similar to daily tasks in a cluttered table. Picking and releasing an object was done by pressing the joystick button. Results not only show the feasibility of our approach and the rapid incremental learning of desired robot motion from a short number of demonstrations, but also that our approach enables the customization of robot trajectories according to the user’s individual preferences.

Results

Electrophysiological signature of ErrPs. As hypothesized, error expectation (i.e., perception of an eventual collision) elicited ErrPs in subjects’ brain. Figure 3a and Supplementary Fig. 2 illustrate the grand average of all the data collected during the first and second experiments over all subjects (N = 13) of the EEG channel FCz, located in the fronto-central midline, for trajectories where the subject released the joystick to avoid a perceived collision (in red, N = 110 ± 32 per subject) and for trajectories where subjects did not feel the urge to stop the robot (in blue, N = 295 ± 25 per subject). This ErrP grand average has been obtained with a causal filter (4th order bandpass Butterworth filter with cut-off frequencies [1, 12] Hz) that is necessary for online real-time analysis of the EEG. Such a causal filter distorts the signal, what explains why the grand average does not resemble the usual waveform of an ErrP. Supplementary Fig. 3 reports the grand-averaged signals with the equivalent non-causal filtering (forward and backward, using the same Butterworth filter as for the causal version) that clearly exhibits the presence of the error-related negativity followed by a positive peak, which corresponds to the typical waveform of ErrPs although with different timing as reported in Fig. 3a. The reason for the appearance of the negative and positive peaks earlier than usual is that, in our case, EEG is synchronized to joystick release and not to the onset of the robot action that makes the subject judge the trajectory as risky. In line with previous works15,21, the sequential negative and positive deflections were observed for the erroneous condition. Also, as shown in the scalp-wide topographical representations (Fig. 3b, top-right and bottom-middle), the first positive peak at 0.15 s and the following negative at 0.5 s were strongly modulated over the fronto-central areas.

A potential confounding factor that may give rise to the EEG potential associated with erroneous trials is that subjects released
**Fig. 1 Overview of the control architecture and experimental protocol.**

- **a** The robot follows trajectories generated from a planar dynamical system. The workspace of the robot (i.e., the table) is modeled with a vector-field and the robot’s trajectories are generated from the position initial position. Therefore, the robot follows a specific vector to reach its target.

- **b** An illustration of our approach. The robot moves towards the cube autonomously avoiding the glass with trajectories generated by a dynamical system. However, some trajectories (red dashed line) pass very close to the glass, creating a feeling of uncertainty to the user as the robot may collide with the glass (i.e., obstacle). This error expectation elicits ErrPs in the brain activity of the user and the output of the ErrPs decoder is associated with the robot trajectories. The desired trajectories are computed with the use of IRL.

- **c** The experimental protocol on the first experiment. The robot moves from left to right and vice versa performing an obstacle avoidance. The dashed dark lines correspond to the random trajectories of the robot, some of them could result in collision with the obstacle. The subject can deflect the joystick right or left to direct the robot accordingly or release the joystick for correcting the motion. This protocol corresponds to the calibration session of the second experiment too.

- **d** The experimental protocol in the second experiment. The subject commands the robot to grasp the object and place it on one of the four target positions (dashed circles) by pushing the joystick left, right, back or forward. The crimson objects correspond to the different obstacles placed in between the target positions. The green dashed line presents the target options for the user.

**Fig. 2 Information flow for training the robot’s controller.** For each demonstration, the output of the ErrP decoder is converted into a weight and introduced to the IRL method together with the demonstration. Then, IRL infers a new trajectory that would lead to high reward on the basis of previous demonstrations. The resulting trajectory is afterward used to configure the DS-modulation parameters ($\rho, \eta$) using gradient descent. Finally, the controller uses these parameters to generate the next robot’s trajectory, based on a dynamical system approach, that should better reflect user’s preferences and guarantee obstacle avoidance.
Fig. 3 ErrP decoding results. a Grand average over all subjects (N = 13) of the EEG channel FCz, located in the fronto-central midline, for erroneous trajectories (N = 110 ± 32 per subject, in red) and for correct trajectories (N = 295 ± 25 per subject, in blue) during the calibration phase of the two experiments. For erroneous trials, time 0 s corresponds to the moment where subjects released the joystick; while for correct trials, the blue trace corresponds to the average EEG potential in the period [1.25, 2.5] s with respect to the onset of the robot motion. Gray area represents the time interval used for building the ErrP decoder. Inset: Topographical representation of EEG amplitude over the subjects scalp for erroneous trials at three different time points with respect to the onset; i.e., 0.00, 0.15, and 0.50 s. b Time-frequency analysis of the FCz channel, grand average over all erroneous trials and subjects. Time 0 s corresponds to the moment where subjects released the joystick. c Classification performance in four different conditions (mean ± std). The red, blue, green, and purple bars represent average time-locked classification accuracy of offline recordings, time-locked classification accuracy of online recordings, and continuous classification accuracy of online recordings, respectively. The vertical dotted line separates subjects who participated in the first experiment (S01-S08) from subjects who did in the second experiment (S09-S13). *p < 0.01. d Latency to detect ErrPs during continuous robot motion in the online phase (N = 73 ± 22 per subject). The box plot represents the distribution of the decoding latency with respect to the moment when subjects released the joystick (time 0 s). Vertical dotted line as in c. e Averaged correct and erroneous robot trajectories obtained during the calibration phase (mean ± std) of the two experiments. Black dashed line indicates the averaged release time, and the rectangles of each color indicate the time window used for computing the decoder (see section "Decoding the error-related potentials" for detail). f Error rate over 10% intervals of the calibration phase of the two experiments. Each dot corresponds to error rate of a subject during the specific time period of calibration, and the box plot illustrates the distribution of error rate for each interval of the calibration phase.
the joystick. Such a change in motor behavior should be accompanied by modulations of the mu and beta rhythms (broadly [8, 30] Hz) over the contralateral motor cortex that could extend over adjacent areas, the fronto-central midline FCz in particular. However, as depicted in Fig. 3b, the main modulations observed in FCz is an increase of the theta rhythm ([4, 8] Hz), which is considered as the main oscillatory pattern of error monitoring²⁴²⁶. We have further addressed the effect of joystick usage in a control experiment that compared the neural correlates of error expectation when participants either used the joystick as in the calibration phase of the two experiments or just monitored the robot trajectories while someone else controlled the joystick (see subsection "Effects of joystick usage on ErrP").

**ErrP-decoder performance.** Figure 3c reports classification accuracy of the ErrP decoder in four different conditions. Supplementary Figs. 4 and 5 report individual confusion matrices of the four conditions for the first and the second experiments, respectively. We collected 61 ± 18 correct and 11 ± 4 erroneous trials in the first, and 88 ± 42 correct and 18 ± 15 erroneous trials in the second experiment per subject. On average across all subjects, 84 ± 8%, 71 ± 9%, 83 ± 11%, and 70 ± 13% (mean ± std) of classification accuracy was obtained for **Offline-Timelock**, **Online-Timelock** and **Online-Continuous** decodings, respectively. Offline performance measures were determined with data from the calibration phase performing a 10-fold cross validation, while Online decodings report classification accuracy achieved during the adaptation phase when performing the classification analysis either in a time-lock or in a continuous manner. As expected, **Timelock** reached the highest performance for Offline and Online conditions as it was evaluated on the time-locked windows used to build the ErrP decoder. This accuracy substantially decreases during the continuous evaluation of the decoder along the robot trajectory in the adaptation phase. Nevertheless, **Online-Continuous** did not further decrease with respect to **Offline-Continuous**. Furthermore, as illustrated in Fig. 3d, despite this lower online performance, the mean latency across all 13 subjects in decoding ErrPs with respect to the moment when subjects released the joystick was 0.76 ± 0.15 s. This value was very close to the latency of the encoding with the Time-Lock approach, 0.5 s. The rationale for this latency of the Timelock approach was because the decoder was trained with the time window of [0.0, 0.5] s with respect to the release of the joystick. Decoding latency was similar between the two experiments, 0.78 ± 0.14 and 0.73 ± 0.18 s, respectively.

In order to assess statistical differences of the classification accuracy across the four different conditions, we performed a two-way repeated-measures ANOVA, the first factor is Offline or Online, while the second factor is Timelock or Continuous. The ANOVA revealed a significant difference between **TimeLock and Continuous** (F(1, 12) = 27.1, p < 0.001), but not between **Offline and Online** (F(1, 12) = 0.55, p = 0.47), nor a significant interaction between the two factors (F(1, 12) = 0.004, p = 0.95). The subsequent post-hoc analysis with Bonferroni’s critical value correction revealed significant differences between **TimeLock and Continuous** in both Offline and Online conditions, but not between other pairs of conditions (Fig. 3d and Table 1).

We further analyzed the relationship between the ErrP classification performance and the release rate during the adaptation phase of the two experiments by performing Pearson’s correlation coefficient analysis for the conditions **Online-Continuous** (r(13) = −0.655, p(13) = 0.015) and **Online-Timelock** (r(13) = −0.770, p(13) = 0.002) (Supplementary Fig. 6). As expected, subjects with lower ErrP classification performance had a higher release rate of the joystick due to inaccurate weighting of the robot trajectories.

**Learning the desired parameters with inverse reinforcement learning.** In the first experiment, we investigated the robot trajectories with or without the release, and the error rate of the calibration phase to understand when and how often participants released the mouse over the course of the calibration phase. As shown in Fig. 3e, participants released the joystick when the robot was moving along a lower trajectory, and so passing closer to the obstacle, compared to the trajectory in which participants did not release the joystick. The rationale for this behavior is that these lower trajectories elicited error expectation. Upon the release, we observed an elbow shape in the trajectory to avoid the obstacle without collision. The number of collisions was 4 ± 1 per subject during the first batch with random modulation parameters, i.e., before the modulation parameters were generated by IRL. Additionally, we analyzed the release rate during the calibration phase to examine if there was a trend in the subjects’ behavior over time. Figure 3f illustrates the error rate for each 10% interval of the calibration phase; a Pearson’s correlation revealed no trend of the release rate across subjects during the course of calibration (r(130) = 0.037, p(130) = 0.67), endorsing no effect of habituation on releasing the joystick.

As shown in the previous subsection, the performance of the ErrP decoder varies among the subjects, and its accuracy decreases during online operation. If the output of the ErrP decoder were used directly to control the assistive robotic arm, its low performance would result in inconsistent and unreliable trajectories. Here, we present a method for dealing with this issue and increasing the efficiency of the robot. Figure 2 illustrates the control approach. To learn the preferred trajectories for each subject, we associate each demonstrated trajectory with a weight coming from the posterior probability of the ErrP decoder (W = 1 – PostProb(ErrP)). Then, we employ an IRL method with these demonstrations to converge to the desired trajectory.

The first step of the IRL consists of an update of the reward function. The reward function is an implicit model of the subject’s costs function. In our implementation, IRL expresses the reward function as a weighted linear combination of Gaussian kernels, which are radial basis functions, centered on the obstacle. Each kernel has a different width. The superposition of the kernels delineates the preferred regions around the obstacle. As learning proceeds, the width for each kernel changes and the region may be enlarged or shrunk to reflect the subject’s preference.

Figure 4a illustrates the evolution of the reward function after three consecutive rounds of IRL adaptation for one of the subjects. As mentioned, updates are influenced by the probability of having decoded an ErrP. The lower the probability, the higher importance is given to the trajectory. This is visible by looking at the first two learning rounds (first two graphs in Fig. 4a). The introduction of a demonstration with a significantly higher weight than the previous one moves the area with high reward (yellow ring) closer to this demonstration. As more demonstrations are introduced, the size of the area with high reward increases (third graph of Fig. 4a), covering the demonstrations of high weight and rejecting the demonstration with lower weight. Thus, the generated trajectory (gray dashed line in Fig. 4a) is closer to the trajectories of high weight.

In practice, the ErrP decoder may assign quite different weights to similar trajectories as depicted in Fig. 4b–l. Although the trajectories (black and red lines) are similar, the ErrP decoder assigns a significantly smaller weight to the red trajectory with respect to the two black trajectories (0.91 and 0.81 to the black trajectories and 0.35 to the red trajectory). Nevertheless, the
The weighted IRL tolerates this inconsistency and generates a trajectory (i.e., green dashed line) in between the trajectories the system considers to be correct. Figure 4d further illustrates how our IRL module handles the variability of the ErrP decoder, which not always assign the correct weights to the trajectories. The ErrP-decoder assigned small weights to 4 out of the 5 first demonstrations (Fig. 4d-i). As a result, the new generated trajectory lies in-between these four demonstrations and the uncorrected demonstration to which the ErrP decoder assigned a weight of 1. For the next 5 robot trajectories (Fig. 4d-ii), the subject applied no correction although the ErrP-decoder assigned different weights to each of the trajectories, indicating differences in the subjective evaluation of their quality. Hence, the IRL-based learning scheme manages to produce a trajectory closer to the preferred trajectory for this subject. Furthermore, although no corrections were applied by the subject, 2 out of the 5 robot trajectories in Fig. 4d-ii had weights less than 0.4, likely indicating cases of false positives. Similar to the previous case, the weighted IRL tolerates the inconsistency of the ErrP-decoder.

Besides the above case of false positives, the variability of the EEG signals could also make the ErrP decoder generate false negatives; i.e., the decoder fails to detect the ErrP associated to a robot trajectory which the subject did correct. Figure 4b-ii shows such a case where the subject applied a correction to the red trajectory that leads to a sharp change of direction. However, the ErrP decoder falsely assigned a large weight (0.84%) to this trajectory, in the same range as the black continuous trajectory for which the subject did not apply any correction. Still, the erroneously classified trajectory had only a minor influence on the new trajectory generated by the IRL module, which remains in close proximity to the other two (correct) trajectories. Thus, the weighted IRL approach exhibits a high tolerance to inconsistencies of the ErrP decoder.

The approach is not only robust to the natural variability and sub-optimal performance of the ErrP decoder, but it also converges rather quickly. Indeed, once the modulation parameters are generated by our learning scheme (i.e., after the first 3 or 5 trajectories, depending on the experiment), the subject did not need to correct the robot motion in the large majority of the ensuing demonstrations, as shown in Fig. 4c. In both experiments, the number of corrections during adaptation is significantly lower than before IRL initialization (i.e., initial trajectories during the calibration phase generated with random modulation parameters). Specifically, a two-sample t-test on the ratio of the number of corrections over the overall trials before and during adaptation returned a p < 0.001 in the first experiment (60 corrections out of 224 trials before adaptation and 9 out of 336 during iterative adaptation, over all eight subjects), regardless of the number of trials used for IRL initialization, and p = 0.0443 in the second experiment (41 corrections out of 192 trials before adaptation and 24 out of 320 during adaptation, over all five subjects). Supplementary Fig. 7 presents the percentage of corrections of each subject. Interestingly, in the first experiment, the subjects corrected the robot trajectories 2.2 ± 1.0% and 2.5 ± 1.3% of the adaptation trials when using 3 and 5 demonstrations for IRL initialization, respectively. A two-sample t-test between the number of corrections when the number of trials used for IRL initialization was 3 or 5 showed no significant differences (p = 0.87). This indicates that 3 demonstrations are efficient for identifying the subject’s preferred trajectories. Furthermore, no significant differences were noticed between the frequency of corrections and the stage of the adaptation phase; the corrections were not concentrated on specific sets of trials (one-sample t-test over the frequency of correction occurrences over all subjects, p = 0.78 for the 3 demonstrations and p = 0.65 for the 5 demonstrations).

We further evaluated this assumption of fast IRL convergence in the second experiment, where the subjects were asked to interact with the robot to perform more complex pick-and-place tasks. The subjects corrected 8.25 ± 3% of the adaptation trials, 3.75 times more than in the first experiment (p = 0.04). The increase in the number of corrections was expected due to the increase on the task complexity, since this protocol not only involved avoiding obstacles, but also picking and moving objects to multiple targets. In addition, the subject’s viewpoint to the robot motion affected the number of corrections. Although no significant differences were noticed on the frequency of corrections among the different robot motions (p = 0.58, one-way ANOVA), the subjects corrected on average 65.9% less the robot motion when the motion direction was perpendicular to the subject’s field of view (perpendicular to the sagittal plane) than sideways. No significant differences were noticed between the frequency of corrections and the stage of the adaptation phase; the corrections were not concentrated on specific sets of trials (p = 0.58, one-sample t-test over the frequency of correction occurrences over all subjects). Moreover, all the subjects drove the robot to all of the four potential targets, for more details see Supplementary Fig. 8.

Furthermore, it is worth noting that our approach enables the customization of robot trajectories according to the subject’s preference. Figure 5a presents the final trajectories of two subjects for the four sets of adaptation trials (i.e., the learned trajectories at the end of each evaluation) in experiment 1 when using 3 initial demonstrations. The trajectories of subject 5 pass closer to the obstacle than those of subject 4 who prefers a more conservative robot behavior. Also, and importantly, the learned trajectories for each subject are consistent. Customization to individual preferences is also depicted on the distribution of the learned DS-modulation parameters (Fig. 5b) for the two subjects, which are different. This is also the case across all subjects as illustrated in Fig. 5c. Supplementary Figs. 9 and 10 provide more details on the learned parameters for all subjects in experiment 1. Although, given the small number of parameters, there is an overlap across subjects, the values of the parameters still result in different personalized trajectories. Moreover, as shown in Fig. 5d, the convergence to individual parameters is not arbitrary, but reflects the actual behavior of the subjects during the calibration phase. Taking as a reference the subjects’ behavior during the calibration phase.

### Table 1: Results of the post-hoc analysis between each pair of conditions, indicating the estimated difference between the corresponding two marginal means of classification performance, the standard error of the estimated difference, and the corresponding p-value.

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Estimated difference [%]</th>
<th>Standard error [%]</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online-Timelock</td>
<td>Offline-Continuous</td>
<td>12.7</td>
<td>1.8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Online-Timelock</td>
<td>Online-Timelock</td>
<td>12.9</td>
<td>3.3</td>
<td>0.02</td>
</tr>
<tr>
<td>Offline-Timelock</td>
<td>Online-Timelock</td>
<td>1.0</td>
<td>1.6</td>
<td>0.55</td>
</tr>
<tr>
<td>Offline-Continuous</td>
<td>Online-Continuous</td>
<td>1.27</td>
<td>1.8</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Statistically significant p-values are in bold.
phase (400 trajectories of which 25% were perceived as erroneous by subjects) used for calibration of the individual ErrP decoders, the KL-divergence is significantly lower between the distributions of the modulation parameters learned by the IRL method and the parameters accepted by the subject during the calibration phase than between the distributions of learned parameters and the parameters corrected by the subject during the calibration phase (two-sample t-test, \( p < 10^{-3} \)). This indicates that the learned parameters correspond indeed to a subset of the parameters the subject considers acceptable.

**Effects of joystick usage on ErrP.** Although the neural correlate of error expectation when participants released the joystick have the properties of an ErrP, the question arises as to whether error expectation without the use of the joystick would have the same neural correlate. To answer this question, and rule out that the neural correlate of error expectation is not elicited by the interaction with the joystick or overlapping components, we performed a control experiment in a setup identical to the calibration phase of the two experiments, where the robot arm moved from one side of the table to the other while avoiding an obstacle.
Supplementary Video 1 presents an overview of our methods and results together with our robotic implementation.

Discussion
We have described and experimentally validated a novel approach for assistive robotic manipulators that people with residual motor capabilities, although lacking fine control, can easily operate and rapidly train to perform desired behaviors. Our approach combines inverse reinforcement learning (IRL) techniques and brain–computer interfaces (BCI) that decode error-related potentials (ErrP), enabling the robotic system to infer a reward function from the subject’s ErrP that leads to individual preferred control policies without requiring the participant to make it explicit. Such a combination avoids the need to collect optimal demonstrations, something that people suffering from severe motor disabilities can hardly do. Instead, the intelligent robotic manipulator automatically generates trajectories that are weighted by the output of the ErrP decoder, indicating whether or not the user considers them appropriate. These weighted trajectories are continuously fed to the IRL module to achieve seamless adaptation to the user’s preferences.

Results not only show the feasibility of our ErrP-IRL approach and the rapid incremental learning of desired robot motion from a short number of demonstrations (Fig. 4), but also that our approach enables the customization of robot trajectories according to the subject’s individual preferences (Fig. 5). Critically, the learned parameters reflect the actual behavior of the subjects during the calibration phase. Furthermore, since our approach requires a small number of training samples, it scales very well, and exhibits good generalization capabilities, to more complex tasks like in our second experiment where the systems learns the subject’s preferred trajectories in 8 different conditions for pick-and-place tasks while avoiding other objects on the table.

Fig. 5 Final parameters and robot’s trajectories after learning. a Example of preferred trajectories for two subjects in experiment 1 (Subjects 4 and 5). Subject 5 prefers trajectories closer to the obstacle. b The distributions of the learned modulation parameters for Subjects 4 and 5. The preferences of the two subjects are depicted on the different distributions of the learned parameters. The distribution of the learned parameters for Subject 4 occupies a higher region than the one of Subject 5. c Map of Hellinger distances among the distributions of the learned modulation parameters between all the subjects of experiment 1. d Comparison of the KL-divergence (mean ± std) between the distributions of the modulation parameters learned by the IRL method against the parameters corrected and the parameters accepted by the subject during the calibration phase (400 trajectories of which ~25% were perceived as erroneous by subjects) used for calibration of the ErrP decoder. *p<10^{-3}. In the middle. In this control experiment participants either used the joystick as before (condition “with-joystick”) or monitored the robot trajectories while an operator utilized the joystick, never releasing it even if the robot arm collided with the object (condition “without-joystick”).

Supplementary Fig. 11 shows the grand-averaged signals of the two classes, i.e., erroneous and correct, in the two conditions, i.e., with-joystick and without-joystick. We collected 232 ± 39 correct and 64 ± 34 erroneous trials in the with-joystick condition, as well as 272 ± 16 correct and 31 ± 15 erroneous trials in the without-joystick condition. We observed a reduced number of erroneous trials in the without-joystick condition as participants indicated their subjective preferences retrospectively, i.e., after completing a trial. The number of collisions between the end-effector and the obstacle was 3 ± 2 times in the with-joystick condition, whereas it was 6 ± 2 times in the without-joystick condition. To confirm whether these time-locked neural responses are significantly different due to the motor action of releasing the joystick, we performed a Wilcoxon’s signed-rank test for each time sample of the signals between the with- and without-joystick conditions for each class, followed by a Benjamini–Hochberg false discovery rate correction\(^{30,31}\). The statistical analysis revealed no significant difference between the two conditions for both the erroneous and correct classes. Furthermore, we computed the Pearson’s correlation coefficient of the two conditions within the time window of [−0.1, 0.4] s, independently for each class (erroneous: \(r(256) = 0.869, p(256) < 0.001\)); correct: \(r(256) = 0.006, p(256) = 0.93\)). These results confirm that the neural correlate of error expectation is elicited by the perception of an erroneous trajectory that leads to a collision and not by the interaction with the joystick or multiple overlapping components.

Supplementary Table 2 presents an overall summary of the experiments and their results. Supplementary Video 1 presents an overall summary of the experiments and their results. Supplementary Table 2 presents an overall summary of the experiments and their results. Supplementary Video 1 presents an overview of our methods and results together with our robotic implementation.
Another key property of our approach is its robustness to the natural variability and sub-optimal performance of the ErrP decoder when deployed online and in a continuous manner.

In contrast to other brain-controlled robotic studies45,53, our approach does not require assistance from a virtual interface for controlling the robot. The increased autonomy of the autonomous robotic system reduces subject’s efforts to control the robot. This is advantageous over other approaches where the subject was required to continuously modulate the brain activity for controlling position57,24 and orientation22 of the end-effector. Correcting discrete erroneous robot actions with ErrPs and RL methods has shown great potentials in robot control55,32,33. Here, we extend the use of ErrPs in a number of ways: (i) they are associated not only to explicit erroneous actions, but also to error expectation; (ii) they can be decoded also during continuous robot motion; and (iii) they carry enough information to learn the subject’s preferred robot trajectories rather than only the motion direction.

In the present study, ErrP were not elicited when participants observed an explicit erroneous robot action, but by an error expectation—i.e., the moment during the robot continuous movement that the user considers will lead to an erroneous trajectory because it will not meet the participant’s preferred obstacle avoidance behavior. The results of the control experiment show that the neural correlate of error expectation is similar, no matter whether subjects use a joystick to interact with the arm robot or not. This neural correlate is an event-related potential whose morphology and topography corresponds to an ErrP, as it consists of the two well-known electrophysiological negative and positive deflection around the fronto-central area of the brain, i.e., error-related negativity (ERN) and error positivity (Pe)35,35. ERN started deflecting even before the release of the joystick (around −0.1 s before the release, see Fig. 3a and Supplementary Fig. 3), and peaked around 0.01 s, a timing considered too early to rely on external sensory feedback30. Pe was observed around 0.3 s after the release, following ERN. It has been suggested that Pe may be a delayed parietal P300 associated with the perception of erroneous actions37,39, a hypothesis supported by the covariation of the Pe amplitude after errors in a Simon task to the P300 amplitude in response to variations in the inter-trial interval40.

Although ErrP has been previously exploited for teaching robots52,53, in these previous works the robot made discrete movements that facilitated the elicitation and decoding of the ErrP. Here we demonstrate the presence of ErrP even during continuous robot motion (Fig. 3), a challenge in the BCI field. Indeed, when exactly the subject considers that the robot motion is erroneous may well be an incremental decision-making process, whereby evidence is accumulated over time31, which varies across trajectories and subjects. To better characterize this decision, our experimental protocol asked subjects to release the joystick used for interaction in order to avoid a perceived collision. Individual ErrP decoders were trained by aligning EEG data to the time of joystick release, and then deployed in a continuous manner during the online trajectory-adaptation phase. Despite the expected decrease in performance, online continuous decoding was stable and did not further degrade with respect to the estimated offline-continuous performance with data collected during the calibration phase. Furthermore, ErrP decoding exhibited a short latency, which was very close to the latency of the optimal decoding with the Time-Lock approach (Fig. 3d).

Previous works have explored the presence of ErrPs during continuous tasks42−48, but in most of them, erroneous events were still generated in a discrete manner such as sudden discrepancies in the execution of commands delivered by the subjects. Grimes et al.47 took a step further and analyzed ErrPs arising from robot trajectories that gradually deviate from targets. Nevertheless, in their experimental protocol, the deviations happened always at the same moment along the trajectories, which facilitated the offline analysis. In our work, the robotic manipulator followed a variety of trajectories, spanning a large workspace. Deviations were variable and happened when subjects determined so, according to risks of collision. More importantly, our work is the first to have demonstrated in an online setting the possibility to decode in real-time the presence of ErrP elicited during continuous robot movements (see also recent work by Lopes-Dias et al.48, where ErrP were also decoded during continuous robot movements, but elicited by robot failures (stop movement) followed by a large perpendicular displacement with respect to its direction of movement).

To learn the preferred adaptation of the robot trajectories according to the subject, we integrated the output of the ErrP-BCI into an IRL scheme. Levine et al.26,27 show that Gaussian-Process IRL (GP-IRL) is able to learn a reward function, representative of the task, even with sub-optimal demonstrations. Deriving from this outcome, we associate the demonstrations with a weight, coming from the ErrP-BCI, which defines the level of optimality of the demonstrations. Different from the original formulation, we assume that we have access to this information, as conveyed by the probabilistic output of the BCI. This allows us to take advantage of both the GP-IRL and the availability of naturally elicited brain signals to modulate the learning. The original GP-IRL employs a modular log-likelihood which makes it vulnerable to the initial random parameters of the optimization. Although this case was rarely noticed in our experiments, a re-initialization is needed to converge to an optimum.

In our experiments, the subject directs the robot using a joystick, which is a frequent practice in controlling robot arms from individuals with motor disabilities49−51. However, it requires some residual distal muscle activity on the upper limb, which makes it inaccessible to subjects with severe motor impairments like paralysis (e.g., high spinal cord injury) or degenerative conditions (e.g., ALS, MS). In our experimental set-up, the joystick serves solely as a reliable indicator of the ground-truth, critical for the evaluation of our approach, and a target definition for the robot. Moreover, our learning scheme for the preferred robot trajectories avoids any input from the joystick, and depends only on the output of the ErrP decoder and the robot demonstrations. Hence, the joystick interface could be replaced with another interface, e.g., a eye-tracker, for target selection, without modifying our control approach. There is evidence that gaze and visual guidance can be used for the selection of robot actions52−54. Since the subject is not required to constantly modulate the brain signals in our approach, using an additional module for target selection based on vision should not increase subject’s fatigue to unacceptable levels.

Generating robot trajectories using dynamical systems (DS) provides the robotic system with the flexibility to rapidly modify the robot trajectories for avoiding the obstacles, whilst guaranteeing the system’s convergence to the goal. Since the trajectory modification depends on two parameters, we exploit this characteristic for relating the output of the ErrP-decoder to the desired robot trajectories. In our work, we modulate an originally linear DS based on these two parameters in order to avoid convex obstacles. Since this method guarantees the stability of the system, our approach could be further expanded to the modulation of non-linear dynamical systems12 and to non-convex obstacles13. However, the obstacle avoidance method modifies the trajectories of the end-effector of the robotic manipulator, whilst the overall configuration of the robotic manipulator depends on an inverse-kinematics solver (e.g., ik-solver). As the ik-solver is agnostic of the obstacle’s position and shape, the solution from the ik-solver might result in the collision of the obstacle with another robot link. In this work, we address this issue by letting the robotic manipulator move above the obstacles, assuming that all the
objects and obstacles are placed on a plane. The introduction of the obstacles’ characteristics into the ik-solver decreases the possibility of collision. The incorporation of these approaches to our system could enable the obstacle avoidance in the joint-level of the robotic manipulation without removing the benefits of trajectory generation of the robot’s end-effector from a DS.

Another limitation of our approach is that, in our current implementation, obstacles were static and located in pre-defined positions. To bring this work closer to a real-world application, the robot should be able to detect the targets and obstacles’ positions. To bring this work closer to a real-world application, trajectory generation of the robot is a crucial aspect.

To avoid obstacles, the initial linear dynamical system can be modulated locally. In the decoder-calibration phase of the second experiment, we increased the number of trials per condition. The subject could move the robot to go left or right with the joystick, whilst the robot was attempting to perform an obstacle avoidance. The subject was allowed to release the joystick and correct the robot trajectories, if they perceived a potential collision (note that IRL will not necessarily lead to learning preferred DS parameters similar to the position of the end-effector and the target’s position). Therefore, joystick release (i.e., DS parameters that increase the distance to objects), since the robot avoided the trajectories and not the corrected DS-modulation parameters.) As IRL requires a batch of demonstrations, the modulation parameters were chosen randomly. Afterward, the DS-modulation parameters for the following ten trials were generated from our weighted IRL method. In this second phase, we evaluated (1) the output of the ErrP decoder and (2) the performance of the control approach with the utilization of IRL. Hence, the trajectory-adaptation phase corresponded to our testing phase. The metric of performance for the learning of the preferred trajectories was the number of trials in which the subjects modified the DS modulation (by releasing the joystick) after the IRL method was initially trained with the first batch of demonstrations. In the first experiment, the trajectory-adaptation phase consisted of eight sets of ten trials. In four of out of the eight sets of trials, the batch size for the IRL was 5 demonstrations and for the remaining four sets of trials the batch size was 3. In the trajectory-adaptation of the second experiment, we increased the complexity of the robot motion bringing it closer to a real application. Specifically, the subject controlled the robot to pick and place objects from/to four positions, releasing the joystick in case of error expectation. We placed different obstacles in between the target positions, letting the robot perform obstacle avoidance autonomously. The subject selected the target position with the joystick and to grasp or release the object by pressing the joystick downwards. Since this experimental setup involved more targets, the IRL method learned trajectories for eight conditions; one for each set of targets times the factor grasping/not-grasping. Of targets and obstacles and the functionality of the joystick. The subject freely directed the robot to move towards one of the four targets and to grasp or release the object by pressing the joystick downwards. Since this experimental setup involved more targets, the IRL method learned trajectories for eight conditions; one for each set of targets times the factor grasping/not-grasping. Of targets and obstacles and the functionality of the joystick. The subject freely directed the robot to move towards one of the four targets and to grasp or release the object. Once a batch of 3 demonstrations for each condition were acquired, using random DS-modulation parameters, the IRL method was trained and produced the preferable modulation for each condition. The subject performed 40 additional trials, where the last five demonstrations were used iteratively to retrain the IRL method for the chosen condition. Subjects repeated the trajectory-adaptation phase of the second experiment twice.

Dynamical system and obstacle avoidance. The robot trajectories are generated from a linear first-order autonomous dynamical system (DS):
In our approach, we define a priori the parameters of the obstacle and let the subject customise the DS-modulation parameters $\rho$ and $\eta$ exploiting the ErrPs elicited from their brain activity.

**Inverse reinforcement learning.** Let us say that a task is defined by continuous states $s = (x_1, \ldots, x_T)$ and continuous actions $a = (a_1, \ldots, a_T)$, such that the next state is a result of the previous state and the corresponding action:

$$f(s_{t-1}, a_t) = s_t$$

(3)

The actions of each state are defined by a policy $\pi$ as a result of a reward function $r(s, a)$; inverse reinforcement learning (IRL) utilizes the observations of optimal behaviors from experts in order to learn a reward function that describes the human actions$^{58}$. The observations correspond to sample paths of an agent that follows an underlying policy $\pi^*$. In this paper, we employ IRL for learning the most preferable robot trajectories according to the subject.

In our approach, the agent corresponds to the end-effector of the robot while the observations are the demonstrations of the robot motion, i.e., the trajectories of the end-effector. Let us define the state as the position of the end-effector $s_1$ and its velocity $s_2$ as the action $u$. Due to the non-linear trajectories of the demonstrations, we employ a Gaussian process (GP) as the reward function for mapping the raw values to rewards, similar to the approach of Levine et al.$^{26,27}$ Specifically, the GP covariance $K_t$ is a variant of the radial basis function (RBF) kernel, so that the corresponding features $F$ of a point are:

$$f = \exp\left(-\frac{1}{2}K \cdot \theta \cdot (\Lambda \cdot \theta)\right)$$

(5)

where $\lambda$ and $\Lambda$ correspond to hyper parameters. The center of the GBP kernel are the center of the obstacle. We put three kernels at each obstacles, with different width $\lambda$ and covariance matrix $\Lambda$. More kernels lead to more computation time and three kernels is sufficient in our experiment.

In order to find the preferred trajectories, we need to maximize the following GP likelihood:

$$L = \log(p(y|\theta, \beta, F)) = -\frac{1}{2}K \cdot \gamma \cdot \gamma \cdot \frac{1}{2}K + \log(\log(\lambda + 1))$$

(6)

where $p(y|\theta, \beta, F)$ corresponds to hyper parameters prior that guarantees the sparsity of $\lambda$ and prevents degeneracies that occur when $\gamma \rightarrow 0$. Following the work of Levine et al.$^{26,27}$, we select this prior to be:

$$\log(p(y|\theta, \beta, F)) = -\log(\log(\lambda + 1)) - \frac{1}{2}K \cdot \gamma \cdot \gamma \cdot \frac{1}{2}K$$

(7)

Once the likelihood is optimized, we can use the reward to retrieve the expert's policy. The reward at a feature point $f(s, u)$ is given by the GP posterior mean:

$$r(s, u) = K'_s \cdot \gamma \cdot \gamma \cdot K'_u$$

(8)

where $K'_s$ is the covariance between $f(s, u)$ and the inducing points. In the case of multiple demonstrations, we maximize the sum of the accumulated likelihood:

$$L_n = \sum L_s$$

(9)

with $n$ being the number of demonstrations introduced to the method.

A fundamental assumption of IRL stands on the optimality of the provided demonstrations. However, this is rarely the case, especially in control methods that require input from neurophysiological signals due to their low signal-to-noise ratio.

In this work, we address this uncertainty by assigning a weight $w_t$ to the provided demonstration in relation to the brain activity of the subject. Thus, the optimized likelihood becomes:

$$L_n = \sum w_t L_s$$

(10)

where $w_t = 1 - \exp(\rho \cdot \eta)$, with $\exp(\rho \cdot \eta)$ being the posterior probability that outputs the ErrP decoder.

Once we learn the reward function from IRL, we compute the corresponding modulation of the DS. Since the modulation is described by a pair of parameters $(\rho, \eta)$, we employ the basic gradient-free gradient descent, using non-linear simplex, to compute the modulation parameters which give the closest reward to the one computed by IRL. The gradients of $\rho$ and $\eta$ for each step $\gamma$ (and $g_1$ and $g_2$ accordingly) are given from the formulas below:

$$g_1 = \frac{\partial r(s, a)}{\partial \rho} = \frac{R_1 - \rho \cdot \eta}{\epsilon}$$

(11)

$$g_2 = \frac{\partial r(s, a)}{\partial \eta} = \frac{R_2 - \rho \cdot \eta + \epsilon}{\epsilon}$$

(12)

where $\gamma$ and $R$ correspond to the learning step (selected to be $10^{-5}$) and the learned reward accordingly. In our experiments, GP-IRL required 10–15 iterations for producing a trajectory, regardless of the number of demonstrations introduced (3 or 5). The computational time required for the generation of a new set of DS-modulation parameters was between 10 and 40 s, 13.2 2.4 and 32.5 8.1 s when the initial batch consisted of 3 demonstrations and 5 demonstrations, respectively.

We further investigate the differences among three types of modulation parameters across subjects: (1) the parameters of the trajectories corrected by the subjects during the calibration phase used to calibrate the individual ErrP decoders, considered as erroneous parameters; (2) the parameters of the trajectories that the subjects did not correct during the calibration phase, considered as correct parameters; and (3) the parameters learned by the IRL method, or learned parameters. For each subject, we model the parameters of each of the above three clusters with Gaussian distributions. We use the Kullback–Leibler divergence (KL-divergence)$^{58}$ for exploring the similarities among the distributions for each subject. KL-divergence is a relative measure of similarity between two distributions and, in our case, it reveals whether the learned parameters are more similar to the correct parameters than the erroneous parameters.

Additionally, we examine whether the IRL method converges to the same parameters for all the subjects or it offers a customized solution for each individual subject. To do so, we use another type of divergence metric, namely the squared Hellinger distance$^{59}$. In contrast to the KL-divergence, the squared Hellinger distance is bounded with values between 0 and 1, where 0 indicates that the means of the distributions are identical, and reaches its maximum value 1 when the distributions do not overlap.

**Decoding the error-related potentials.** We recorded 16 EEG and 3 electro-oculogram (EOG) signals at 512 Hz via two USBamps (g.tec medical technologies, Austria). EEG electrodes were located at Fz, FC3, FC1, FCz, FC2, FC4, C1, C3, Cz, C4, C2, O1, OZ, P3, Pz, P4, CP3, P1, CP1, CPz, CP2, and CP4 (10-20 international system), while the 3 EOG electrodes were placed at above the nasion and below the outer canthi of the eyes. The ground electrode was placed on the forehead (A10) and the reference on the left earlobe. The EEG and EOG signals were notch filtered at 50 Hz to eliminate the power noise. To reduce signal contamination, participants were asked to restrict eye movements and blinks during experiments.

Before experiments, participants underwent 90 s of recording in which they were asked to perform clockwise and counter-clockwise rolling of eyeballs, vertical and horizontal eye movements and repeated eye blinks. This data was subsequently used to compute coefficients to linearly subtract EOG artifacts from EEG signals based on the autocovariance matrix of EEG and EOG signals$^{58}$.

We removed trials in which subjects mis-operated the joystick. Specifically, for each subject, an erroneous trial was considered anomalous and removed if it was associated with a joystick release time that deviated from the mean reaction time for that subject more than a pre-defined threshold, i.e., mean-absolute deviation (MAD) > 3.

**Time-frequency analysis.** Main components of the ErrPs were identified by performing a time-frequency analysis with the Stockwell transform$^{58}$. S-transforms are performed on the data recorded during the decoder-calibration phase of the first and second experiments. To remove baseline drift, we firstly applied a 2nd order causal Butterworth high-pass filter with the cut-off frequency of 1 Hz. The reason for employing a causal filter is to emulate the signal processing conditions of online decoding analysis. Then, EEG signals were epoched in the time window $[-0.5, 1.0]$ s with respect to the onset of the joystick release in erroneous trials, or $[1.0, 2.5]$ s with respect to the onset of the robot movement in correct trials. S-transform $S_s(t, f)$ of EEG signals $s(t)$ is defined as follows:

$$W_z(t, d) = \int_0^\infty x(t)w(t - \tau, d)dt$$

(13)

$$S_s(t, f) = \int_0^\infty \exp\left(-\frac{t^2}{2}\right)\exp(-2\pi i ft)dt$$

(14)

$$w(t, f) = \frac{\Im \{W_z(t, d)\}}{\sqrt{2\pi}} \exp(-2\pi if t)$$

(15)

$$S_s(t, f) = \int_0^\infty \exp\left(-\frac{t^2}{2}\right)\exp(-2\pi if t)dt$$

(16)

We used different time windows for erroneous and correct trials because, while joystick release is a natural marker for erroneous trials, no such marker exists for correct trials. For these correct trials, we opted to follow a common technique, which is to arbitrarily choose a segment of the EEG signals$^{58,59}$. We employed the time window $[1.5, 2.0]$ s with respect to the onset of the robot movement, as this is the period before the robot passes around the obstacle and where the subjects should closely monitor the situation, i.e., the robot, the obstacle, and the distance between them.
The final step for the time-frequency analysis was to compute the event-related spectral perturbation (ERSP) of the ErrP potentials at the beginning of the trajectory, and keep the consistency with the time baseline time window was to avoid the presence of visually-evoked event-related potentials at the beginning of the trajectory, and keep the consistency with the time window of correct trials in subsequent classification analysis. The rationale for the baseline time window was to avoid the presence of visually-evoked event-related potentials at the time window used to create the ErrP classifier (see below). The range of time and frequency was set to $[0.5, 10]$ s and $[1, 30]$ Hz, respectively. Since the ErrP is known to appear over the frontal-central area, this analysis was performed on the EEG signal at the FCz channel (Fig. 3b).

**ErrP decoding.** During the BRL trajectory-adaptation phase of the first and second experiments, we decoded the presence of ErrPs online to weight robot trajectories in a seamless fashion. To build the individually customized ErrP decoder, which infers the presence or absence of ErrPs in a continuous manner, we used the data from each decoding trial, i.e., the sliding window calibration Butterworth bandpass filter with the cut-off frequencies [1 12] Hz. Then, EEG signals were segmented into epochs for ErrP trials, we selected the window [0.5 10] s, whereas for correct trials, the window was [1.5 2.5] s with respect to the onset of the robot movement, which corresponds to the start of a trial. Figure 3e shows these time windows as well as the grand average of robot trajectories for ErrP and correct trials during the calibration phase of the first and second experiments.

To enhance the signal-to-noise ratio (SNR) of the ErrPs signals, we applied a spatial feature extraction approach analysis. This spatial filter method transforms the averaged ErrPs to a subspace containing different ERP components. Only the first three components were kept for further analysis. For each ErrP condition, we extracted three different type of features: the EEG voltage per time sample after downsampling the data to 32 Hz (48 temporal features); power spectral densities (PSD) per EEG component (15 PSD features); and features on the Riemannian manifold, which computes a low-dimensionality manifold representation from a non-linear combination of the EEG component space (21 covariance features); thus, 84 features in total. In order to extract Riemannian features, we concatenated the EE...
signals of each class, i.e., erroneous and correct, between the two conditions based on Winkler’s signed-rank test followed by the Benjamini–Hochberg false discovery rate correction. Additionally, we computed the similarity of the grand-averaged signals between the two conditions with Pearson’s correlation analysis independently for the two classes.

Data availability

The source data for graphs and charts are available as Supplementary Data. All physiological and empirical data is available at https://zenodo.org/record/3627015.

Code availability

Custom code to implement obstacle avoidance movement is available at https://github.com/mlpd/lascr_1.5b_obstacle_avoidance. Physiological analysis was performed by using MATLAB R2018b using the BioSig toolbox (http://biosig.sourceforge.net). Code for stockwell transform is available at https://www.mathworks.com/matlabcentral/fileexchange/51086-time-frequency-distribution-of-a-signal-using-a-transform

Received: 10 January 2020; Accepted: 10 November 2021;

Published online: 16 December 2021

References


Acknowledgements
This work was funded by the Hauser Foundation, Switzerland, and by the Swiss National Centre of Competence in Research in Robotics (NCCR). We would like to thank Dr. Bastien Orset for his assistance in the control experiment.

Author contributions
I.B., F.I., S.W., R.C., J.d.R.M., and A.B. conceived and designed the experimental protocols. I.B., F.I., and S.W. implemented the experimental protocols. I.B., F.I., S.W., and C.G.P.R.C. were responsible for data acquisition. F.I. designed and implemented the EEG decoder and performed the analysis. F.I. and J.d.R.M. interpreted EEG results. I.B. and S.W. designed and implemented the IRL method and the robot controller. I.B. performed the analysis and interpretation of robot trajectories and IRL results. I.B., F.I., S.W., R.C., J.d.R.M., and A.B. prepared the draft manuscript. All authors reviewed the results and approved the final version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-021-02891-8.

Correspondence
Andreas Bülthoff, aude.billard@gmail.com

Peer review information Communications Biology thanks Elia Kirchner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Karli Montague-Cardoso.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access
This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021
2 Research Output of the Natural Language Processing Group
2 Research Output of the Natural Language Processing Group

The NLP research team, led by Prof. Cieliebak, develops technologies for the analysis, understanding and generation of speech and text. We combine methods from linguistics, natural language processing (NLP) and artificial intelligence to enable natural language communication between humans and machines.

In our research, we work on topics such as text classification (e.g., sentiment analysis, hate speech detection), chatbots/dialogue systems, text summarization, speech-to-text, speaker recognition and natural language generation. At the end of 2021, the group consists of 10 researchers: 1 full professor, 4 senior scientists, 4 research assistants, and 1 IAESTE intern. Two of the research assistants pursue their master’s degree, and two additional master’s students are not employed at CAI. In October 2021, Jan Deriu successfully defended his PhD.

One major topic within the research team is speech processing for Swiss German, for example within a joint SNF project with FHNW and the University of Zurich, to build speech-to-text systems for Swiss German dialects. Here we are closely collaborating with the Swiss Association for Natural Language Processing (SwissNLP) for collecting new audio data in Swiss German; and we are supervising several bachelor projects around this topic. We are proud that one of the bachelor theses, by Bogumila Dubel, was awarded with the “Siemens Excellence Award 2021”.

Another focus topic is summarization of dialogues. In the on-going project “Interscriber” (funded by Inno-suisse), we have already developed a solution for automatic transcription of interviews and meetings. The next step is now to summarize such meetings automatically. We started with an extensive literature study that resulted in a survey paper on the field and are now exploring various methods to solve this challenging question.

In “Virtual Kids”, a long-term SNF project together with HSLU, we are developing a simulator for children in police interrogations. In this dialogue system, police officers learn how to properly interview children who are victim or witness of a crime. The major challenge is to model a child with a “hidden story” (what really happened) and how it behaves in such interrogations. Finally, we are working intensely on detecting misuse of social media, i.e., recognition of hate speech, fake news, fake accounts etc. This endeavor is funded by armasuisse and the ZHAW Digital Futures Fund with the goal to gain a better understanding of how misuse affects society and how it can be prevented.

We would like to thank all project partners, collaborators, funding agencies and students for their support!

The NLP 2021 team

Mark Cieliebak, Don Tuggener, Manuela Hürlimann, Jan Milan Deriu, Nikolaos Kapotis, Pius von Däni-ken, Katsiaryna Mlynchyk, Nicola Good, Daniel Neururer, and Zhivar Sourati
Towards Understanding Lifelong Learning for Dialogue Systems

Mark Cieliebak and Olivier Galibert and Jan Deriu

Abstract Lifelong learning is the ability of a software system to adapt to new situations during its lifetime. We explore how this paradigm can be applied to dialogue systems, how it might be implemented, and how we can evaluate the lifelong learning progress.

1 Introduction

Chatbots, dialogue systems, conversational user interfaces - the names may differ, but the basic idea is the same: “intelligent” computer systems that can interact with humans in natural language. These systems have become more and more popular in the past years, and there is an increasing interest in spoken and written dialogue systems in research and industry. Prominent examples include automatic customer support agents, smart home devices such as Amazon Alexa or Apple’s Siri, and in-car operating systems. While implementing a successful and reliable dialogue systems is already a challenge, ”lifelong learning” even adds an additional twist: the system should be able to adapt to new situations during its lifetime. More precisely, the dialogue system learns to handle new situations by interacting with its environment (e.g. asking a domain expert, scraping the web), instead of being retrained by a machine learning expert. For instance, a chatbot for travel advice might be confronted with a new location that is not yet in its knowledge base. One strategy to deal with this situation could b to ask the user to give additional information (e.g.
in which country, GPS coordinates etc.), then explore the web to find information about the location (e.g. databases, Wikipedia or travel reviews), and finally analyze, structure, and integrate the information into the chatbots’ knowledge base.

In this paper, we attempt to make a step forward towards understanding what lifelong learning in the context of dialogue systems means. In order to achieve this, we first briefly introduce both concepts independently and discuss typical settings and applications. Then we describe the the impact of applying lifelong learning to dialogue system (in Section 4). Finally, we turn to the important question how we could measure the success (or failure) of lifelong learning in the context of dialogue systems (Section 5).

2 What is a Dialogue System?

In the following, we introduce the concept of a dialogue system. A dialogue system allows the user to converse with a computer system using natural language. Such systems can be applied to a variety of tasks, e.g.:

- **Virtual Assistants**, which are developed to aid its users in every-day tasks, such as scheduling appointments. They usually operate on predefined actions, which can be triggered by voice command.
- **Interaction with Information Systems**, by asking questions or finding a piece of information (e.g. the most suitable hotel in town).
- **Training environments**, where the dialogue systems are developed to train students in the interaction with medical patients or train military personnel in questioning a witness.
- **Answering Questions**, where the dialogue system can answer specific questions of a user. These might be factoid questions or more complex questions.

Dialogue systems usually structure dialogues in **turns**, each turn is defined by one or more **utterances** from one speaker. Two consecutive turns between two different speakers is called an **exchange**. Multiple exchanges constitute a **dialogue**. Another correlated view, is to interpret each turn or each utterance as an action (more on this later). The main component of a dialogue system is the dialogue strategy, which defines the content of the next utterance and thus the behaviour of the dialogue system. There are many different approaches to design a dialogue strategy, which are partly dictated by the application of the dialogue system. However, there are three broad classes of dialogue systems, which we encounter in the literature: task-oriented systems, conversational agents and interactive question answering systems. We identified the following characteristic features, which help differentiate between the three different classes: is the system developed to solve a task, does the dialogue follow a structure, is the domain restricted or is it open domain, does the dialogue span over

---

1 In recent literature, the distinction is made only between the first two classes of dialogue systems [1, 2, 3]. However, interactive question answering systems cannot be completely placed in either of the two categories.
multiple turns, are the dialogues rather long or efficient, who takes the initiative, and what is the interface used (text, speech, multi-modal).

- **Task-oriented systems** are developed to help the user solve a specific task as efficiently as possible. The dialogues are characterized by following a clearly defined structure, which is derived from the domain. The dialogues follow mixed initiative: both, the user and the system can take the lead. Usually, the systems found in the literature are built for speech input and output. However, task-oriented systems in the domain of assistance are built on multi-modal input and output.

- **Conversational agents** display a more unstructured conversation, as their purpose is to have open-domain dialogues with no specific task to solve. Most of these systems are built to emulate social interactions and thus longer dialogues are desired.

- **Question Answering (QA) systems** are built for the specific task of answering questions. The dialogues are not defined by a structure as with task-oriented systems, however they mostly follow the question and answer style pattern. QA systems may be built for a specific domain, but also be tilted towards more open domain questions. Usually, the domain is dictated by the underlying data, e.g. knowledge bases or text snippets from forums. Traditional QA systems work on a singe-turn interaction, however, there exist systems that allow multiple turns to cover follow-up questions. The initiative is mostly done by the user who asks questions.

### 3 What is Lifelong Learning?

In the most abstract way, *Lifelong Learning (LL) is the ability of a system to use past experiences to adapt to future challenge.* There exist various definitions of LL in the literature, for instance in ?? For the purpose of this paper, we exploit the definition of LL from Chen and Liu [4], which we summarize in the following:

Lifelong learning is a continuous learning process. Given that the learner has learned $N$ tasks. When faced with the $(N+1)$th task the learner leverages past knowledge to help learn the new task. The goal is to optimize on both the new task and the previous tasks. The three components are: continuous learning, knowledge accumulation and maintenance and leverage past knowledge to learn new tasks. There are some additional considerations to be made considering the above definition.

- The learner learns new tasks continuously, however, in contrast to transfer-learning, the learner improves or at least does not deteriorate its performance on the old tasks. Ideally, by learning new tasks, the performance on the previous tasks improves.

- The learner is not restricted to a certain task or domain. On the contrary, the learner is encouraged to learn different types of tasks (e.g. sentiment analysis, named entity recognition, etc) and domains.
Ideally, the learner is self-motivated and able to find its own learning tasks and data by interacting with the environment.

Note that this definition is strongly focused on “knowledge improvement”, whereas in the setting of LL for dialogue systems, there are also other goals, as we discuss in the following section. In addition, we would like to mention that the concept of a ”task” may cover situations of varying complexity, ranging from single new instances (e.g. a new person in a face recognition system) up to new domains (e.g. switching from cooking to car tuning for a questions answering systems). Finally, note that several other terms have been coined and used for very similar learning paradigms of systems that improve over time, such as continuous learning [5, 6], meta-learning [7], active learning [8], or transfer learning [9]. For a more elaborate introduction of LL, see the recent book by Chen and Liu [4], which gives a good overview of LL in general and describes applications in various fields.

4 Lifelong Learning for Dialogue Systems

While the definition of lifelong learning given by [4] is very general, we attempt not to apply the definition to dialogue systems. These systems allow its users to converse with a computer via natural language. This implies a high level of interactivity. Thus, the focus of applying LL to dialogue systems should lie in the interactive nature of the dialogue. Furthermore, LL describes the capability of the dialogue system to learn to handle new situations throughout its deployment, i.e. without being re-trained by a machine-learning expert. Ideally, the learning takes place in a self-driven and autonomous manner. This does not exclude (it rather encourages) the assistance of a ”domain expert”, i.e. a type of user who takes the role of a teacher.

We assume that the dialogue system is an agent that interacts with its environment. The environment includes humans as well as having access to structured and unstructured knowledge sources (e.g. knowledge bases, Wikipedia, Twitter). When faced with a new situation, the dialogue system has to learn how to handle this new situation. This does not necessarily means that the dialogue system directly adapts to the new situations. Rather, through interaction with its environment, the dialogue system learns to handle the situations over time.

There are various aspects to a dialogue system which can be improved over time:

- **Language Understanding**: Here, the dialogue systems’ capability of parsing the user input is the focal point. This is the case, for instance, when new request types occur over time, for instance if a system was only faced with simple factual questions until now, and the system is suddenly confronted with complex questions.

- **Dialogue Behaviour** is concerned with the ”soft” quality factors of a dialogue, such as human-likeness, appropriateness of responses, efficiency of reaching a goal, engaging utterances etc. Typically, the DS asks after a user interaction for feedback, which is then used to improve the behaviour over time. Thus, the DS
leverages past experiences to improve its future behaviour. Note that in this case, there are no external catalysts that trigger LL, but there is an "intrinsic motivation" of the system itself.

- **Knowledge Induction** is concerned with accumulating more information. This means adapting the knowledge base (KB) with new or updated knowledge, which can be factual knowledge or unstructured. Here, new situations are in the context of handling new entities and relations which are not in the current knowledge base.

- **Capability Improvement** is concerned with extending the functionality of the DS. This can range from domain adaption (e.g. moving from asian recipes to the pasta domain) up to integrating new skills (e.g. reporting weather forecasts for a personal assistant).

In each case the dialogue system needs to improve its aspects over time. Each time it is faced with a new situation one or more of the aforementioned aspects need to be adapted. In the context of dialogue systems, this adoption can be done by means of interacting with a "domain" expert. More precisely, the goal is to remove the need to rely on a dialogue system expert who would retrain the different components of the dialogue system. Rather the domain-expert teaches the dialogue system how to handle a new situation through interaction. Note that in some cases the system may be able to learn how to handle the situation autonomously, especially in the case where it can aggregate data from some external sources. Thus, a LL enhanced dialogue system is able to learn to adapt to new situations by interacting with its environment and not by means of retraining components.

## 5 Evaluation of Lifelong Learning for Dialogue Systems

The above definition of LL for dialogue systems sets a strong focus on learning to handle new situations by interacting with its environment. Thus, the LL component of the dialogue system needs to be trained and evaluated with this in mind. More precisely, the interaction with an environment lies at the centre of the training and evaluation. The environment should enable the interaction the dialogue system will encounter during deployment. This includes the interaction with a domain expert.

In general, LL evaluation methods need to be reproducible in order to measure improvements and changes over time. One straightforward way of doing this is to deploy a dialogue system and let humans interact with it. However, this is very time consuming and expensive, and alternatives with less or no humans in the loop are desired. One major issue that is particular for evaluating dialogue systems is that they produce their "result" - the dialogue - during the interaction with their environment. Thus, any automated environment environment has to provide artificial users, and building them can be as complex as building the dialogue system itself.

When it comes to LL evaluation, additional complexity arises due to the fact that the interaction with the expert needs to be simulated as well. For instance, the dialogue system may ask an expert for advice about a new entity or topic. In general,
the evaluation system cannot know in advance which questions the dialogue system will ask - hence, it is hard to simulate.

**Experimental Evaluation Environment**

We are currently working on an experimental setting to automate the evaluation of knowledge acquisition and capability improvement. We work in the cooking domain, where the dialogue system is developed to assist the user by answering questions about cooking, e.g. providing recipes, giving advice or providing tutorials. Typical question might be "How do i prepare linguini, which is answered with a corresponding recipe from the database.

In order to evaluate the LL capabilities of the dialogue system, we deploy it in a simulated environment, which consists of:

- **Evaluation agent**: provides the questions and evaluates the answers given by the dialogue system. The agent institutes new situations by asking about entities which were not present in any training set of the dialogue system (e.g. enchilada), by asking types of questions which the dialogue system did not encounter yet (e.g. "How do i clean my oven?"), or by asking questions about unseen domains (e.g. Chinese food).

- **Expert**: provides advice to the dialogue system when stuck. The dialogue system can ask clarifying questions to the expert before it tries to answer the question of the evaluation agent. However, this comes at a cost, i.e. each interaction with the expert has its fee, and thus, the system should learn to efficiently interact with the expert.

We envision that the dialogue system asks questions from a list of predefined templates, which the (automated) expert can easily parse and answer. These are, for instance, "What is <X>?" or "Is <X> a relevant entity for this question?". The domain expert has at its hand a large collection of pre-recorded dialogues on the domain, and returns extracts of these dialogue that match to the clarification question.

The evaluation measures the capability of the LL component to adapt to the new situations. This capability is measures by the number of interactions needed with the expert system before answering the initial question correctly. A system with a strong LL component should adapt to new situations quickly.

### 6 Conclusion

Implementing lifelong learning for a dialogue system may aim at 1. extending the underlying knowledge base (Knowledge Induction); 2. handling more complex user interactions (Language Understanding); 3. improving the perceived quality of the
resulting dialogues (Dialogue Behaviour); or 4. extending the functionality of the system (Capability Improvement) over time.

While stating these goals is simple, implementing a system that achieves any of these four goals is far from trivial. To the best of our knowledge, most approaches in research currently tackle the first dimension (Knowledge Induction), while there is almost no solution (yet) for the other three.

One important challenge is to evaluate the progress of LL in such systems. In order to avoid time-consuming and costly human evaluations, automated environments are required. We are currently working on such a system, which shall be presented as shared task in 2020.

7 Acknowledgements

This paper has been partially funded by the LIHLITH project, supported by ERA-NET CHIST-ERA and Swiss National Science Foundation (20CH21_174237).

References

The Sentence End and Punctuation Prediction in NLG Text (SEPP-NLG) Shared Task 2021

Don Tuggener  Ahmad Aghaebrahimian
Zurich University of Applied Sciences (ZHAW)
Winterthur, Switzerland
{tuge, agha}@zhaw.ch

Abstract

This paper describes the first Sentence End and Punctuation Prediction in Natural Language Generation (SEPP-NLG) shared task held at the SwissText conference 2021. The goal of the shared task was to develop solutions for the identification of sentence boundaries and the insertion of punctuation marks into texts produced by NLG systems. The data and submissions, and the codebase for the shared tasks are publicly available.

1 Introduction

Sentence End Detection, also known as Sentence boundary disambiguation (SBD) or boundary detection, is the Natural Language Processing (NLP) task of recognizing where a sentence begins and ends. A period is the most common end of sentence indicator in written English as well as many other Indo-European languages. However, a period may be used in a decimal point, an abbreviation, an email address, or other possible cases as well which makes sentence boundary detection a challenge. Other forms of punctuation such as question and exclamation marks, semicolons, comma, etc. add to this challenge. Although sentence boundary detection is considered an almost solved issue for formal written language (Walker et al., 2001), it poses a challenge in terms of meaning distortion and readability in synthetic or automatically translated or transcribed texts such as the output of Automatic Speech Recognition (ASR) or Machine Translation (MT) systems. The punctuation marks in such synthetic text may be displaced for several reasons. Detecting the end of a sentence and placing an appropriate punctuation mark improves the quality of such texts not only by preserving the original meaning but also by enhancing their readability.

The goal of the SEPP-NLG shared task is to build models for identifying the end of a sentence by detecting an appropriate position for putting an appropriate punctuation mark.

2 Related Work

Similar to the system proposed by Grefenstette and Tapanainen (1997), the earliest attempts for sentence boundary detection utilize a set of rules or regular expressions. In a different direction, Reynar and Ratnaparkhi (1997), and Kiss and Strunk (2006) proposed an information-centric approach based on the Maximum Entropy model, and an unsupervised method based on collocation statistics respectively. Decision tree classifier (Riley, 1989), Naive Bayes (López and Pardo, 2015) and deep learning based (Kaur and Singh, 2019) models are the most recent advances based on machine learning that are proposed for predicting correct positions for the period in particular and other punctuation marks in general. Moving forward and combining the rule-based and machine learning-based systems, Deepamala and Ramakanth (2012) proposed a hybrid system with high performance.

Our task is closely related to Tilk and Alumäe (2016) and follow-up work that uses the Europarl and TED talk corpora for punctuation prediction. Similar to our goal, Żelasko et al. (2018); Donabauer et al. (2021) investigate sentence boundary detection in unpunctuated ASR outputs of spo-
ken dialogues based on textual features. Cho et al. (2017) propose a method to predict sentence boundaries and punctuation insertion in a real-time spoken language translation tool. In a similar setting, Klejch et al. (2017) include acoustic features to improve punctuation prediction in a speech translation system, and Yi and Tao (2019) combine lexical and speech features for punctuation prediction in a traditional ASR setting. Finally, Rehbein et al. (2020) investigate the annotation and detection of sentence like units in spoken language transcripts.

3 Task Overview

Ultimately, the goal of SEPP-NLG is to predict sentence ends and punctuation in NLG texts. However, there are no corpora that feature NLG texts and their manually transcribed and corrected versions. Therefore, we approximate the setting by using a) transcripts of spoken texts, and b) lower-casing the texts and removing all punctuation marks. While there are multiple corpora of transcribed spoken language, we choose the Europarl corpus (Koehn, 2005) as the source for our data. The Europarl corpus consists of transcripts of the sessions of the European parliament and features transcripts in multiple languages.

We offer the following subtasks:

- **Subtask 1**: (fully unpunctuated sentences-full stop detection): Given the textual content of an utterance where all punctuation marks are removed, correctly detect the end of sentences by placing a full stop in appropriate positions.

- **Subtask 2**: (fully unpunctuated sentences-full punctuation marks): Given the textual content of an utterance where all punctuation marks are removed, correctly predict all punctuation marks.

Participants were free to choose for which languages and subtasks they contributed a submission, but were encouraged to participate in all languages.

3.1 Data

We leverage the open parallel corpus (OPUS) version of the Europarl corpus (Tiedemann, 2012) for extracting the task data as it provides sentence boundaries and tokenization. Albeit the sentence boundaries in the corpus are automatically generated, they are quite reliable as the data and the models trained to detect the boundaries contain all the original punctuation symbols of the transcripts.

In the spirit of the “Swissness” of the SwissText conference where SEPP-NLG 2021 is co-located, we select 3 of the 4 official languages of Switzerland, i.e. German, French, and Italian and complement the selection by incorporating English.

The Europarl corpus contains multiple punctuation symbols. For subtask 2, we gauged which subset of them represents a realistic and feasible goal for their automatic prediction in a stream of unpunctuated, lower-cased tokens. Also, we considered which punctuation marks improve the readability of a text the most. Hence, we consolidated the selection of punctuation symbols for subtask 2 to : ; . ! (0 indicating no punctuation) and mapped the symbols ; ; to ; . The period. We removed all sentences from the data that contain other punctuation symbols such as parentheses, as there is no straightforward way to remove punctuation without interfering with the naturalness of a sentence. This removal affected the data for both subtasks and resulted in removing less than 10% of the data per language. We also removed HTML artifacts, and special (non-visible) characters (zero width space, soft hyphen) from the data. Finally, we omitted sentences with fewer than 3 tokens and documents with fewer than 2 sentences.

The data format is as follows: Lower-cased tokens per file are listed vertically, and the labels for subtask 1 (binary classification) and 2 (multiclass classification) are appended horizontally, separated by tab. The labels encode whether a token emits a sentence end (subtask 1) and a punctuation symbol (subtask 2). Table 1 shows an example.

Per language, we randomly selected 80% of the documents for the training and 20% for the test set. From the the training set, we then randomly sampled 20% of the documents as the development set.

Table 2 shows several statistics of our data. We see similar properties for all languages: Most sentences are unique, and there are few sentences that occur both in the train and test sets. German fea-
tutes the largest vocabulary, as is expected due to its morphological richness, and the vocabulary overlap between train and test sets is roughly 50% for all languages.

Concerning the labels, the data is highly skewed towards the 0 label for both tasks, as most tokens do not emit a sentence end or punctuation symbol after them. For example, there are 9,618,776 tokens with the label 0 and 420,446 with label 1 subtask one in the English test set, which yields an average sentence length of almost 24 tokens. Table 3 shows a breakdown of the label counts in the English test set for subtask 2. It shows that the period and comma symbols have similar counts and are the most frequent labels among the non-0 labels. The remaining labels occur less than an order of magnitude less frequently. These label distribution properties are similar across all languages.

### 3.2 Surprise Test Data

The Europarl corpus covers domain-specific language, i.e. political statements in the European parliament. To measure how well the participating systems trained on our data generalize to out-of-domain data, we incorporated a surprise test set comprised of TED talk transcripts (Reimers and Gurevych, 2020).

For each language, we sampled 500 TED talks, favoring those that have the lowest vocabulary overlap with our Europarl test sets to maximize the vocabulary shift. The document-based average percentage of the vocabulary overlap ranges from 85 to 90, meaning there are on average 10-15% of tokens per document in the surprise test set that are not in the Europarl test set.

While being one order of magnitude smaller than the Europarl test set, the surprise test set is also highly and similarly imbalanced regarding the label distribution. In the English surprise test set, there are 67,446 tokens with label 1 and 1,014,464 tokens with label 0. This yields an average sentence length of 16 tokens, which is significantly lower than the 24 tokens in the English Europarl test set. The label counts for subtask 2 follow an almost identical distribution in both test sets.

### 4 Submissions

**ZHAW-mbert:** We provided a baseline based on the multilingual BERT model (Devlin et al., 2019), mBERT, implemented in the simpletransformers library. We treat the task as a token classification problem and segment the documents into subsequent, non-overlapping chunks of length 512 to adhere to the sequence length restrictions of BERT. We fine-tuned the model on the training data of all languages with a randomly shuffled file order across all languages and vanilla settings for about one week on a single GPU.

**ZHAW-adapter-mbert:** To contrast the resource-intensive fine-tuning of mBERT with a computationally cheaper approach of task adaption, we apply the adapter-transformers library (Pfeiffer et al., 2020). Instead of updating all the weights of the base models (mBERT in our case), the adapters approach inserts a few feed-forward layer in between the transformer blocks and only trains those for adapting a base model to a new task. We again use the vanilla settings and train the model for one day.

**OnPoint:** In their study of sentence segmentation, Michail et al. (2021) proposed a majority-voting ensemble model consisting of several Transformer models trained in different ways. The models’ predictions are leveraged at test time using a sliding window to obtain the final predictions. They offered their system as language-dependent models for all four languages of the shared task and both sub-tasks.

---

10 https://github.com/ThilinaRajapakse/simpletransformers
11 https://github.com/Adapter-Hub/adapter-transformers/
Table 2: Training data statistics, showing number of (unique) sentences and tokens and the number of sentences and tokens in both training and test set (\textit{train} \cap \textit{test}) per language.

<table>
<thead>
<tr>
<th>Lang</th>
<th>#sentences unique \textit{train} \cap \textit{test}</th>
<th>#tokens unique \textit{train} \cap \textit{test}</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>1'406'577</td>
<td>33'779'095</td>
</tr>
<tr>
<td>DE</td>
<td>1'308'508</td>
<td>28'645'112</td>
</tr>
<tr>
<td>FR</td>
<td>1'236'504</td>
<td>32'690'367</td>
</tr>
<tr>
<td>IT</td>
<td>1'132'554</td>
<td>28'167'993</td>
</tr>
</tbody>
</table>

Unbabel-INESC-ID: Rei et al. (2021) extend the architecture proposed by Rei et al. (2020) to develop a multilingual model for sentence end and punctuation prediction. Their system is designed based on pre-trained contextual embeddings and built on top of a pre-trained Transformer-based encoder model. They propose their method as a single multilingual model for all languages and subtasks of the shared task.

UR-mSBD: Donabauer and Kruschwitz (2021) propose a system based on a pre-trained BERT model and fine-tuned for the first sub-task. They use language-specific models for each of the four languages of the shard task. They consider sub-task 1 as a binary classification problem by identifying tokens that indicate the position of a full stop.

oneNLP: Applying multi-task Albert for English and multi-lingual Bert for other languages Mujadia et al. (2021) explored the impact of using contextual language models for sentence end and punctuation prediction. They modeled the problem in both subtasks as a sequence labeling task. They presented the results of employing a baseline CRF, as well as the results of applying a fine-tuning method over contextual embedding.

HULAT-UC3M: Based on the Punctuator framework (Tilk and Alumüe, 2016) which is a bidirectional recurrent neural network model equipped with an attention mechanism, Masiello-Ruiz et al. (2021) developed an automatic punctuation system named HULAT-UC3M. They trained HULAT-UC3M for all languages as well as both sub-tasks in the shared task individually.

HTW: Guhr et al. (2021) modeled the task as a token-wise prediction and examined several language models based on the transformer architecture. They trained two separate models for the two tasks and submitted their results for all four languages of the shared task. They advocated transfer learning for solving the task and showed that the multilingual transformer models yielded better results than monolingual models. By pruning the Bert layers, they also showed that their model retains 99% of its performance without 1/4 of the last layers.

5 Results

In section 3.1 we showed that our data is highly imbalanced regarding the label distribution. Accuracy or Macro F1 scores are not suitable metrics in this setting, as majority class prediction would yield an accuracy of 96% for subtask 1 on the English test set, e.g. Therefore, we applied the following metrics to evaluate the participants’ submissions:

- **Subtask 1**: F1 score of the label 1 (the positive class, i.e. sentence end)
- **Subtask 2**: Macro F1 of the selected punctuation symbols

We observe that a) most systems achieve a very high score for subtask 1 for all languages on the Europarl data, and b) the F1 scores are almost identical (with seemingly minor differences in precision and recall) for the top-ranking systems for both tasks. Further, the top-ranking systems are the same ones for both tasks. This is to be expected to some degree, as it can be argued that subtask 2 subsumes subtask 1.

While the F1 scores for subtask 2 seem low compared to subtask 1, a more detailed results analysis reveals that the lower (Macro) F1 scores mainly stem from the labels with the lowest counts in the data. Table 6 gives the detailed classification report.
for the top three ranking system for the English test set. It shows that the systems are able to predict periods, commas, and question marks reliably, but that they struggle with hyphens and colons, which lowers the Macro F1 scores.

<table>
<thead>
<tr>
<th>Label</th>
<th>htw+t2k</th>
<th>OnPoint</th>
<th>Unbabel</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>.</td>
<td>0.82</td>
<td>0.82</td>
<td>0.80</td>
</tr>
<tr>
<td>,</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>:</td>
<td>0.42</td>
<td>0.41</td>
<td>0.37</td>
</tr>
<tr>
<td>?</td>
<td>0.57</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>0.88</td>
<td>0.91</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Table 6: F1 scores per label for the top-performing systems on the English test set for subtask 2.

All systems perform significantly worse on the surprise test sets for both tasks. To gauge the difficulty of the task on the TED dataset compared to the Europarl dataset, we train the ZHAW-mbert approach on the remaining TED talks that were not selected for the surprise test set and then test the system on the surprise test set. Table 7 shows that the average F1 score does improve by 11 percentage points when training the ZHAW-mbert system on domain data. Still, the 0.66 F1 score is 9 percentage points behind the average F1 score on the Europarl data. Hence, the drop in performance of Europarl-trained ZHAW-mbert on the surprise test set can both be accounted for by the domain shift and by the increased difficulty of the target domain (TED talks). We expect that this applies for the performance drop of all systems.

<table>
<thead>
<tr>
<th>Label</th>
<th>htw+t2k</th>
<th>OnPoint</th>
<th>Unbabel</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>.</td>
<td>0.82</td>
<td>0.82</td>
<td>0.80</td>
</tr>
<tr>
<td>,</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>:</td>
<td>0.42</td>
<td>0.41</td>
<td>0.37</td>
</tr>
<tr>
<td>?</td>
<td>0.57</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>0.88</td>
<td>0.91</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Table 7: Results of training ZHAW-mbert on TED talks for subtask 2 (averaged over all languages).
We expected some submissions to use linguistic features such as part-of-speech tags or partial syntax parse trees and hypothesized that such systems would fare better on out-of-domain data. However, all participating systems applied neural encodings of the surface tokens and did not encode linguistic features explicitly. Still, the ranking of the systems remains intact on the surprise test sets.

The top three systems in both tasks all use transformers-based approaches and tackle the tasks in a similar manner. We hypothesize that this is the main reason for near identical performance of the systems in terms of F1 scores. Based on the task results, these three systems seem to produce near-identical output. To better gauge their similarities and differences, we evaluate their outputs for subtask 2 in a pair-wise manner on the English test set. We apply the evaluation metric such that one system output takes the role of the ground truth and the other the one of the system prediction, which yields the F1 scores per class that we leverage as an indicator of the similarity or agreement of the per-token predictions. Table 8 shows the results. While the macro F1 scores and even the per-class F1 scores in Table 6 are highly similar, there are significant differences in this analysis. For example, for the hyphen class, the systems have different predictions in over 30% of the cases, and for colon in roughly 20%. For the majority classes of the non-0 classes, the systems disagree in about 10% of the cases for comma, but their predictions are highly similar for period (96% agreement).

Table 8: System prediction similarity between the three top-performing systems on the English test set for subtask 2.

Following Tuggener (2017), we can take the comparison a step further and analyse the type of differences per label. For example, the OnPoint submission’s F1 score for hyphen is 4 percentage points higher than the one of Unbabel, and their prediction agreement for hyphen is 68%. This does not indicate, however, whether OnPoint’s predictions are always better. The aforementioned comparison takes a ground truth label \( G \), the predicted label \( A \) of one system, and the predicted label \( B \) of another system and defines three types of differences for the cases where \( A \neq B \):

- correction: \( G = B \)
- new error: \( G = A \)
- changed error: \( G \neq A \neq B \)

Table 9 shows the results. We see that the predictions of commas makes up a large portion of the differences. When OnPoint’s prediction differs from Unbabel’s for comma, OnPoint is correct and Unbabel incorrect in nearly 70% of the cases, which explains the 2 percentage point higher performance of OnPoint in Table 6. Still, Unbabel is correct in almost 30% of the cases where the two predictions differ.

Table 9: Detailed comparison of the differences in Unbabel’s predictions versus OnPoint’s predictions for English in subtask 2. \#Diff. signifies the number of tokens that have the respective label as the ground truth and for which OnPoint’s and Unbabel’s predictions differ. The remaining columns represent the percentage of this number in each difference class.

<table>
<thead>
<tr>
<th></th>
<th>#Diff.</th>
<th>corr.</th>
<th>new err.</th>
<th>changed err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>45'552</td>
<td>34.22%</td>
<td>62.59%</td>
<td>3.19%</td>
</tr>
<tr>
<td>-</td>
<td>50'496</td>
<td>69.01%</td>
<td>28.30%</td>
<td>2.69%</td>
</tr>
<tr>
<td>-</td>
<td>16'190</td>
<td>49.28%</td>
<td>44.69%</td>
<td>6.03%</td>
</tr>
<tr>
<td>-</td>
<td>4'422</td>
<td>51.15%</td>
<td>33.04%</td>
<td>15.81%</td>
</tr>
<tr>
<td>-</td>
<td>2'014</td>
<td>41.46%</td>
<td>31.43%</td>
<td>27.11%</td>
</tr>
<tr>
<td>-</td>
<td>1'158</td>
<td>63.90%</td>
<td>29.53%</td>
<td>6.56%</td>
</tr>
</tbody>
</table>

Table 9: Detailed comparison of the differences in Unbabel’s predictions versus OnPoint’s predictions for English in subtask 2. #Diff. signifies the number of tokens that have the respective label as the ground truth and for which OnPoint’s and Unbabel’s predictions differ. The remaining columns represent the percentage of this number in each difference class.

In conclusion, we observe that while the top three systems perform similarly in terms of Macro F1 scores for subtask 2, there are nuances to each system that distinguishes them from the others.

5.1 Winners

While we showed that there are differences in the outputs of the top three systems that are not reflected in the averaged F1 scores, the declared criteria for winning the task are the averaged F1 scores in Tables 4 and 5. Since the top three systems in these tables are practically indistinguishable based on these F+ scores, we declare OnPoint, htw+t2k, and Unbabel as the joint winners of the SEPP-NLG 2021 shared task. Congratulations!
6 Conclusions

We presented the setting and results of the first Sentence End and Punctuation Prediction in NLG text (SEPP-NLG 2021) shared task. We found that all participants explored neural networks-based models (particularly transformers) to tackle the task. The results for the in-domain Europarl data were high for the most common punctuation symbols, but the performance decreased significantly when the models were faced with out-of-domain data.

The discussion of the task results during the session at the SwissText conference yielded the following desiderata for future iterations of the shared task:

- More heterogeneous data (more domains)
- Add truecasing as an additional task
- Add other language families
- Take inference time / computational costs as an additional evaluation criteria, or create a separate track that puts emphasis on a low-resource/low-latency setting

Acknowledgments

We thank the participants for their submissions and their valuable feedback on early versions of the data and task details. This work was funded by Innosuisse under grant project nr. 43446.1 IP-ICT.

References


Ottokar Tilk and Tanel Alumäe. 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In INTERSPEECH.


ZHAW-CAI: Ensemble Method for Swiss German Speech to Standard German Text

Malgorzata Anna Ulasik, Manuela Hürlimann, Bogumila Dubel, Yves Kaufmann, Silas Rudolf, Jan Deriu, Katsiaryna Mlynchyk, Hans-Peter Hutter, and Mark Cieliebak

Centre for Artificial Intelligence
Zurich University of Applied Sciences
{ulas, hueu, deri, mlyn, huhp, ciel}@zhaw.ch
bodubel@gmail.com, y.kaufmann@yagan.ch, silasrudolf@gmail.com

Abstract

This paper presents the contribution of ZHAW-CAI to the Shared Task “Swiss German Speech to Standard German Text” at the SwissText 2021 conference. Our approach combines three models based on the Fairseq, Jasper and Wav2vec architectures trained on multilingual, German and Swiss German data. We applied an ensembling algorithm on the predictions of the three models in order to retrieve the most reliable candidate out of the provided translations for each spoken utterance. With the ensembling output, we achieved a BLEU score of 39.39 on the private test set, which gave us the third place out of four contributors in the competition.

1 Introduction

Speech-to-Text (STT) enables transcribing spoken utterances into text. For successfully performing a transformation from speech to a text, the existence of a standardised writing system of the target language is of prime importance. This is where Swiss German poses a substantial challenge: it does not have a standardised orthography since it functions as the default spoken language in both formal and informal situations, while for writing, the Standard German language is used. This phenomenon, called “medial diglossia” (Siebenhaar and Wyler, 1997), occurs in the entire German-speaking part of Switzerland, which is additionally characterised by a high dialect diversity. Swiss German is increasingly used for writing in informal contexts, but since there is no single standard writing system, Swiss German speakers usually write phonetically in their local dialect in informal situations (Siebenhaar, 2003). On formal occasions such as work meetings and political debate, speech is typically transcribed into Standard German. As there is a considerable linguistic distance between Swiss German dialects and Standard German, developing a model for transcribing Swiss German speech into Standard German text actually involves Speech Translation, which combines STT with Machine Translation (MT) (Bérard et al., 2016).

As a response to the Shared Task “Swiss German Speech to Standard German Text” organised at SwissText 2021, we provided a solution consisting of three models based on different architectures: Fairseq (Wang et al., 2020a), Jasper (Li et al., 2019) and Wav2vec XLSR-5 (Baevski et al., 2020) which were trained with various data sets, both in Standard German and Swiss German. Their predictions were subsequently fed into a majority voting algorithm with the aim to select the most reliable translation.

The remainder of this paper is structured as follows: Section 2 provides the description of the Shared Task and Section 3 discusses relevant literature. In Section 4 we present the systems which make up our final solution, their architecture and the applied training data. In section 5 we provide an overview of all experiments performed with these models and their outputs. Section 6 lays out the ensembling approach and section 7 presents the post-processing experiments we performed on the predictions of the models. The paper ends with a conclusion presented in section 8.

2 Shared Task Description

The goal of the Shared Task was to build a system for translating speech in any Swiss German dialect into Standard German text (Plüss et al., 2021).

The organisers provided a labelled data set con-
taining 293 hours of audio recordings, mostly in the Bernese dialect, transcribed in Standard German. Since the alignment between the recordings and the transcripts was done automatically, each utterance has an Intersection over Union (IoU) score reflecting its alignment quality. Additionally, there was an unlabelled data set consisting of 1208 hours of recordings, mostly in the Zurich dialect. The dialect distribution of the test set is close to the actual Swiss German dialect distribution in Switzerland.

The translation accuracy of the provided solutions is measured using BLEU, a standard metric for automatic evaluation of machine translation (Papineni et al., 2002). The approach consists in counting n-grams in the candidate translation matching n-grams in the reference translation without taking the word order into account. The metric ranges from 0 to 100. A perfect match results in a score of 100. A score of 0 occurs if there are no matches. The tool used by the organisers for evaluating solutions is the NLTK implementation of the BLEU score with default parameters. Prior to evaluation, both the references and the translations are normalised: the utterances are lowercased, the punctuation is removed, the numbers are spelled out and all non-ASCII characters except for the letters “¨a”, “¨o”, “¨u” are removed.

The test set was split into a public and a private subset of equal sizes. For all evaluations presented in this paper, the public test set was used.

3 Related Work

Speech Translation (ST) is the task of translating spoken text in a source language to text or speech in a target language. The approaches to solve this problem can be put into two categories: cascading approaches and end-to-end approaches (Sperber and Paulik, 2020).

Cascaded Approaches work by splitting the task into two steps: first, an STT model transcribes speech of the source language to text in the target language, and then a machine translation (MT) module translates the generated text into the target language (Waibel et al., 1991). The main issue with the cascaded approach is the fact that errors made by the STT module are propagated to the MT module (Ney, 1999). Thus, efforts are put into coupling the STT and MT modules to prevent error propagation, for instance, by generating multiple hypotheses of the STT system via n-best search or the creation of lattices (Woszczyna et al., 1993; Schultz et al., 2004).

End-to-End Approaches model ST as a single task, where input is speech in the source language, and the output consists of text or speech in the target language. The main issue with this modelling approach is the lack of sufficient training data. Whereas data for STT typically consists of several hundreds of hours of transcribed data, most ST datasets contain only a fraction of this amount. For instance, the Europarl-ST corpus contains on average only 42 hours of transcribed data per language pair (Iranzo-Sánchez et al., 2020), whereas the Librispeech STT corpus contains around 1000 hours of transcribed data (Panayotov et al., 2015). For this reason, end-to-end approaches nowadays rely on leveraging multi-task learning and single language pre-training of the STT and MT submodules and use the ST dataset for fine-tuning (Wang et al., 2020).

Most cascading approaches rely on data where access to both the source language transcript and its target language translation is needed. However, in our scenario, we do not have access to written text of the source language since Swiss German is a spoken language, and thus, often directly transcribed into Standard German (see 1 for more details). Thus, our models follow the End-to-End approach.

4 Systems Description

This section describes the architecture of the three models which build the foundation for the experiments presented in Section 5 and are components of the final solution which combines the three models’ outputs in an ensembling algorithm. The section also explains what data was used for training the models.

4.1 Fairseq

4.1.1 Model

Fairseq is based on the transformer architecture for Speech-to-Text provided by Fairseq S2T Toolkit (Wang et al., 2020a), which combines the tasks of STT and ST under the same encoder-decoder
architecture (Changhan Wang, 2020). The experiments were trained with the small transformer model with 256 dimensions, 12 Layers encoder, 6 Layers decoder, 27M parameters, Adam optimiser, and inverse square root for the learning rate scheduler. Decoding is executed with a character-based SentencePiece model (Taku Kudo, 2018) using an n-best decoding strategy with n=5. The acoustic model (encoder) can be pre-trained with the same transformer architecture as described above.

4.1.2 Data
The audios were extracted to 80-dimensional log mel-scale filterbank features (windows with 25 ms size and 10 ms shift) and saved in NumPy format for the training. To alleviate overfitting, speech data transforms SpecAugment (Park et al., 2019), adopted by Fairseq S2T, were applied. For text normalisation we used the script provided by the task organisers. Additional numbers were spelled out using num2words3. We use three additional datasets:

- SwissDial (Pelin Dogan-Schönberger, 2021): 26 hours of Swiss German
- ArchiMob (Tanja Samardzic, 2016): 80 hours of Swiss German
- Common Voice German v4: 483 hours of German4

The SwissDial dataset consists of 26 hours of audios in 8 different Swiss dialects with corresponding transcriptions in Swiss dialect and Standard German translations. The Swiss German transcription rules differ between dialects. ArchiMob contains 70 hours of audios in 14 different Swiss dialects with transcription in Swiss German, where each word is additionally provided with a Standard German normalisation. The transcription rules are normalised and are equal for all dialects (Dieth transcription, (Dieth and Schmid-Cadalbert, 1986)). Common Voice German v4 consists of 483 hours of audios in Standard German with corresponding transcriptions.

4.2 Jasper
4.2.1 Model
We used the Jasper (Li et al., 2019) configuration corresponding to our best submission in the pre-decessor of this Shared Task (Büchi et al., 2020). The Acoustic Model as per Büchi et al. (2020) consists of 10x5 blocks and was pre-trained on 537 hours of Standard German data (see Büchi et al. (2020), Table 2). In all reported experiments, we fine-tuned five blocks on the Shared Task data as described in Section 5.2 below. We used last year’s extended language model, a 6-gram model trained with KenLM, without further fine-tuning on this year’s data. For the data sources, see Table 2 in Büchi et al. (2020). Decoding was done using beam search with a beam size of 1024.

4.2.2 Data
We extracted the audios to 64-dimensional mel-filterbank features with 20ms window size and 10ms overlap as input to the Jasper acoustic model. The reference texts were preprocessed as described in Büchi et al. (2020). No additional Swiss German audio data was used for training Jasper.

4.3 Wav2vec XLSR-53
4.3.1 Model
Wav2vec XLSR-53 is a cross-lingual extension of wav2vec 2.0 as per Baevski et al. (2020). Pre-trained on 53 different languages, it attempts to learn a quantisation of the latent representations shared across languages by solving a contrastive task over masked speech representations. In the experiment below, we fine-tuned wav2vec XLSR-53 on the Shared Task data. No explicit language model was used to conduct the experiment.

4.3.2 Data
The labelled data used for fine-tuning XLSR-53 was based on the task training data. However, it was further pre-processed removing all utterances which contained special characters or were detected as not being in German using langdetect5. Numeric values were replaced by strings using num2words6.

5 Experiments on Individual Models
Sections 5.1 through 5.3 present the experiments we performed to improve the individual models and provide the BLEU scores achieved in each experiment. We also discuss approaches to improve the model outputs with the use of ensembling (Section 6) and post-processing (Section 7).

3https://pypi.org/project/num2words/
4https://commonvoice.mozilla.org/en/datasets/
5https://github.com/Mimino666/langdetect
6https://pypi.org/project/num2words/
5.1 Fairseq

Below we describe the different models and experimental results obtained with Fairseq. All experiments are trained with the same configuration as described in Section 4.1 and can be divided into three groups: extension of training data, inclusion of a pre-trained encoder and ensembling.

5.1.1 Extending the training data

**Fairseq F-SP-0.9** For F-SP-0.9 we trained the model from scratch on the Shared Task training data. We used 176 hours, corresponding to an Intersection over Union (IoU) greater or equal to 0.9.

**Fairseq F-SP-All** We noted that the model F-SP-0.9 generalises very poorly, so for F-SP-All we trained a new model with the entire task training data, which corresponds to 293 hours. Despite partially poorly aligned translations, the model benefits from the new data: the BLEU score is improved by about 4.32 points.

**Fairseq F-SP-SD** We decided to extend the training data with the SwissDial Corpus. For this, we trained a new model F-SP-SD with the entire task training data plus all data from SwissDial. This data extension improves the score by an additional 4.81 BLEU points in comparison to F-SP-All.

5.1.2 Including pre-trained encoder

**Fairseq F-SP-DE** We also investigated how to improve the encoder (acoustic model). We pre-trained a Standard German (DE) encoder on the Common Voice German v4 dataset. For F-SP-DE, we added the pre-trained encoder and trained the model on the entire Shared Task training data. Including the DE encoder improves the score by 3.36 BLEU points in comparison to F-SP-All.

**Fairseq F-SP-SD-DE** Since both models F-SP-SD and F-SP-DE improved the BLEU score, we decided to bring the two approaches together. We trained a new model F-SP-SD-DE with the entire Shared Task training data, SwissDial data and include the pre-trained DE encoder in the training. This brings an improvement of 8.37 BLEU points in comparison to F-SP-All.

**Fairseq F-SP-AM-DE** In this model we used the entire task training data plus the data from Archimob. For the training we included the pre-trained DE encoder. This setup improves the BLEU score by 14.01 in comparison to F-SP-All.

**Fairseq F-SP-SD-CH** In order to further improve the acoustic model, we trained an encoder in Swiss German (CH) on the SwissDial and Archimob dataset. We trained a new model F-SP-SD-CH with the entire Shared Task training data and SwissDial and included the pre-trained CH encoder in the training. The BLEU score in comparison to F-SP-All is improved by 12.54 points.

5.1.3 Ensembling

**Fairseq Ensemble F-SP-SD & F-SP-DE (F-E1)** In this experiment, we ensembled the models F-SP-SD and F-SP-DE. F-E1 achieves a BLEU score of 28.74. Ensembling is done with the implementation provided by the Fairseq S2T Toolkit. In comparison to F-SP-SD-DE, which combines in the training setup the same training dataset SwissDial as F-SP-SD and the same DE encoder as F-SP-DE, the ensembling performs slightly better. In comparison to F-SP-All the BLEU score improves by 9.94 points.

**Fairseq Ensemble F-SP-AM-DE & F-SP-SD-CH (F-E2)** After the good performance of F-E1, we decided to ensemble F-SP-AM-DE and F-SP-SD-CH. This ensembling improves the BLEU score in comparison to F-SP-All by 17.00 points.

**Fairseq F-E2 extended (F-E3)** Finally, we trained a model on the entire available data for Swiss German (task, SwissDial and Archimob) and used this model to perform ensembling on top of F-E2. For time reasons, we were not able to complete the training and the output of this model could not be included in the final solution presented in 6. We only evaluated an intermediate status of the model and achieved a score of 36.83 BLEU points. In comparison to F-SP-All, it improves the score by 18.03 points.

Table 1 shows the public BLEU scores obtained with the Fairseq models on the Shared Task public part of the test set. The table contains additional information about applied train sets and encoders. F-E3 achieved the best performance with a BLEU score of 36.83 on the public part of the test set (37.4 on the private part). In addition to ensembling, the inclusion of a CH encoder in

\[^{7}\text{http://github.com/pytorch/fairseq/issues/223}\]
the training process as well as the extension of the training data with the ArchiMob corpus benefited the model performance most.

Table 1: Fairseq results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Train set</th>
<th>Encoder BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-SP-0.9</td>
<td>task 0.9</td>
<td>training 14.48</td>
</tr>
<tr>
<td>F-SP-All</td>
<td>task all</td>
<td>training 18.8</td>
</tr>
<tr>
<td>F-SP-SD</td>
<td>task, SwissDial</td>
<td>training 23.61</td>
</tr>
<tr>
<td>F-SP-DE</td>
<td>task</td>
<td>DE 22.16</td>
</tr>
<tr>
<td>F-SP-SD-DE</td>
<td>task, SwissDial</td>
<td>DE 27.17</td>
</tr>
<tr>
<td>F-SP-AM-DE</td>
<td>task, ArchiMob</td>
<td>DE 32.81</td>
</tr>
<tr>
<td>F-SP-SD-CH</td>
<td>task</td>
<td>CH 31.34</td>
</tr>
<tr>
<td>F-E1</td>
<td></td>
<td>- 28.74</td>
</tr>
<tr>
<td>F-E2</td>
<td></td>
<td>- 35.80</td>
</tr>
<tr>
<td>F-E3</td>
<td></td>
<td>- 36.83</td>
</tr>
</tbody>
</table>

5.2 Jasper

Below we describe the different models and experimental results obtained with Jasper.

Jasper-FT For Jasper-FT we fine-tune the pre-trained Standard German model on the Shared Task training data. We used 169 hours, sampled from the set with an IoU greater or equal to 0.9, which were augmented to 507 hours using 90% and 110% speed perturbation as in Büchi et al. (2020).

Jasper-PL We noted that the task test set differs acoustically from the training data since different dialects are present and the audio quality tends to be lower. This motivated the creation of Jasper-PL, where we used pseudo-labeling on the test set. More precisely, we used the hypotheses of Jasper-FT on the task test set to fine-tune Jasper-FT for 20 additional epochs.

Jasper-PL-E We decided to further work on the (comparatively) low-quality audio of the task test set and used the Dolby Media Enhance API v1.1 to create a "enhanced" version of the task test set. The Enhance API automatically improves the quality of audio files, e.g. by correcting the volume and reducing noise and hum. We then fine-tuned Jasper-FT on this data, this time using the hypotheses provided by Jasper-PL as labels since these achieve a higher BLEU score.

Table 2 shows the public BLEU scores obtained with the Jasper models on the two different test sets (Jasper-PL-E was only evaluated on the enhanced test set). The best-performing Jasper model is Jasper-PL with a BLEU score of 32.97 on the public part of the test set. Using the enhanced audio data does not confer any advantage on either prediction or pseudo-label fine-tuning compared to the as-is data. We can, however, see the benefit of rather naive pseudo-labelling in this setting where training and testing data are quite different. Future work could expand on the use of pseudo-labelling by using more advanced setups, such as confidence-based (Kahn et al., 2020) or iterative (Xu et al., 2020) pseudo-labelling.

Table 2: Jasper results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test set</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jasper-FT</td>
<td>task</td>
<td>30.8</td>
</tr>
<tr>
<td>Jasper-FT</td>
<td>enhanced</td>
<td>26.4</td>
</tr>
<tr>
<td>Jasper-PL</td>
<td>task</td>
<td><strong>32.97</strong></td>
</tr>
<tr>
<td>Jasper-PL</td>
<td>enhanced</td>
<td>31.92</td>
</tr>
<tr>
<td>Jasper-PL-E</td>
<td>enhanced</td>
<td>32.92</td>
</tr>
</tbody>
</table>

5.3 Wav2vec XLSR-53

Below we describe the model and experimental results obtained with wav2vec XLSR-53.

Wav2vec XLSR-53 FT For wav2vec XLSR-53 FT we fine-tuned the pre-trained baseline (as published on HuggingFace) on the Shared Task training data. We used 227 hours, corresponding to an IoU greater or equal than 0.8. The data was pre-processed as outlined in Section 4.3.2.

Table 3: wav2vec XLSR-53 result.

<table>
<thead>
<tr>
<th>Model</th>
<th>Train set</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>wav2vec XLSR-53 FT</td>
<td>task 0.8</td>
<td>30.39</td>
</tr>
</tbody>
</table>

6 Ensembling

Having trained and evaluated the three models described in Sections 4.1, 4.2 and 4.3, we performed experiments with two ensembling methods: majority voting and a hybrid technique combining majority voting with perplexity calculation. We used the outputs of the best-performing models of each of the three systems, aiming to select the
most reliable translation for each utterance from among them. The best-performing models were F-E2 (BLEU score of 35.80), Jasper-PL (BLEU score of 32.97) and wav2vec XLSR-53 FT (BLEU score of 30.4).

The models were first categorised based on their BLEU scores into a primary, first auxiliary and second auxiliary models. F-E2 with the highest score was selected as the primary model, Jasper-PL with the second best score was set as the first auxiliary model and wav2vec XLSR-53 FT was used as the second auxiliary model.

In the first step, we aligned the hypotheses of the three models and extracted text passages where all three hypotheses agree, leaving only text excerpts where the hypotheses disagree.

**Majority Voting (MV)** The majority voting consisted in collecting votes for each text excerpt defined in the previous step: a particular hypothesis receives a vote for each word it has in common with any other hypothesis. The hypothesis with the most votes is chosen as the best candidate translation. If multiple hypotheses score the same, the output of the model categorised higher in the hierarchy (primary, first auxiliary, second auxiliary) is selected.

**Hybrid Ensembling (HE)** The hybrid ensembling method combines majority voting with perplexity calculation. If more than one hypothesis scores maximum and the hypotheses with the maximum score are not equal, the perplexity of the hypotheses is calculated. To this end, we extended the particular text excerpt with 3 context words preceding and following the excerpt. For these text segments, we calculated perplexity with a pre-trained uncased German BERT model. The hypothesis with the lower perplexity was selected.

The results of the experiments are presented in Table 4. Out of the two algorithms we applied on the data, better results could be achieved with the majority voting. The BLEU score improved by 2.9 points from 35.80 to 38.70 when compared to the result of the best model (F-E2).

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-E2</td>
<td>35.80</td>
</tr>
<tr>
<td>Jasper-PL</td>
<td>32.97</td>
</tr>
<tr>
<td>wav2vec XLSR-53 FT</td>
<td>30.39</td>
</tr>
</tbody>
</table>

### 7 Transcript Post-processing

Next to the Language Models for Speech Recognition, we evaluated an approach to using text-only data by training a supervised ”spelling correction” (SC) model to correct the errors made by the STT model explicitly. Instead of predicting the likelihood of emitting a word based on the surrounding context, the SC model only needs to identify likely errors in the STT model output and propose alternatives. Intuitively, this task highly depends on the baseline model’s quality: if the model transcribes very well, this task can be reduced to simply copying the input transcript directly to the output.

Most recent approaches for transcript post-processing use a transformer-based method: (Liao et al., 2021) use a modified RoBERTa structure and show an increase of 17.53 BLEU points on the self-augmented English Conversational Telephone Speech data set. On the LibriSpeech dataset, (Hrinchuk et al., 2019) show promising results using a pre-trained BERT as initialisation for their spell correction model, while (Guo et al., 2019) take a different approach with a bidirectional LSTM.

We compared different Transformer architectures with their corresponding open-sourced pre-trained models and other post-processing methods. The objective for all transformer models was set to next-sentence prediction (sequence to sequence generation) with a vocabulary size of 30’000, batch size of 16, and beam search set to 5. The models were initialised with pre-trained German embeddings and fine-tuned for up to 120’000 steps on the Shared Task training set described in 2.

- **BERT** (Devlin et al., 2018), having both encoder and decoder initialised with pre-trained weights.
- **DistilBERT** (Sanh et al., 2020), the lightweight alternative to BERT, reducing the training time up to 60%.

---

10 F-E3 as a last-minute submission could not be used for ensembling
11 https://github.com/dbmdz/berts#german-bert

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Voting (MV)</td>
<td>38.70</td>
</tr>
<tr>
<td>Hybrid Ensembling (HE)</td>
<td>37.62</td>
</tr>
</tbody>
</table>

Table 4: Ensembling results. The BLEU score achieved by each model separately and the BLEU score resulting from applying ensembling methods on the models’ outputs (Majority Voting and Hybrid Ensembling)
• ELECTRA (Clark et al., 2020), which uses a more sample-efficient pre-training approach for the encoder, called replaced token detection.

• SymSpell (Garbe, 2020), which is a spelling correction algorithm for correcting spelling errors based on Damerau-Levenshtein distances, stored in a pre-trained dictionary.

The following table shows the BLEU scores on the public test set, when performing post-processing on the output of the majority voting algorithm as described in 6. The Baseline refers to the BLEU score of the non-processed majority voting output.

Table 5: Post-processing BLEU scores on the public test set

<table>
<thead>
<tr>
<th>System</th>
<th>Baseline</th>
<th>Post-processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>38.70</td>
<td>23.26</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>38.70</td>
<td>26.66</td>
</tr>
<tr>
<td>ELECTRA</td>
<td>38.70</td>
<td>14.77</td>
</tr>
<tr>
<td>SymSpell</td>
<td>38.70</td>
<td><strong>30.65</strong></td>
</tr>
</tbody>
</table>

As the evaluations show, most post-processing attempts decrease the overall BLEU score, with SymSpell as the most straightforward approach performing best. Compared with previous work in this area, this could be explained by the limited amount of data available for training the transformer models. Due to lack of performance, we exclude the post-processing step in our final solution.

8 Conclusion

In this paper, we presented our contribution to the Shared Task "Swiss German Speech to Standard German Text" at SwissText 2021. Our solution combines the outputs of three models based on Fairseq, Jasper and Wav2vec XLSR-53 architectures. Because of time and resource constraints, we used only the labeled data set. Out of the 21 experiments we performed with the models, including transcript post-processing and ensembling, we achieved the best result by applying an ensembling method on the outputs of Fairseq model F–E2 (BLEU score of 35.80) as the primary model, and Jasper-PL (32.97) and wav2vec XLSR-53 FT (30.39) as auxiliary models. We processed the three models’ predictions with a majority voting algorithm and this way retrieved the most reliable candidate out of the provided translations for each utterance in the public test set. With this solution, we achieved a BLEU score of 39.39 on the private test set, which resulted in the third place out of four contributors in the competition.

Swiss German is a low-resource language, which makes training an STT or a Speech Translation system a challenging task. However, our experiments show that applying ensembling both on various models of the same architecture (as in Fairseq models F–E1, F–E2 and F–E3) and on models based on various architectures (as implemented in our final solution) trained with limited data can lead to a score improvement of several BLEU points. Pseudo-labeling is another approach which contributes to model enhancement as we could observe with the Jasper-PL model. We will be further investigating these two methods aiming at improving the results despite the limited data currently available for Swiss German.

References


Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.


Are We Summarizing the Right Way? A Survey of Dialogue Summarization Data Sets

Don Tuggener¹, Margot Mieskes², Jan Deriu¹, and Mark Cieliebak¹

¹Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland, [tuge, deri, ciel]@zhaw.ch
²University of Applied Sciences, Darmstadt, Germany, margot.mieskes@h-da.de

Abstract

Dialogue summarization is a long-standing task in the field of NLP, and several data sets with dialogues and associated human-written summaries of different styles exist. However, it is unclear for which type of dialogue which type of summary is most appropriate. For this reason, we apply a linguistic model of dialogue types to derive matching summary items and NLP tasks. This allows us to map existing dialogue summarization data sets into this model and identify gaps and potential directions for future work. As part of this process, we also provide an extensive overview of existing dialogue summarization data sets.

1 Introduction

Dialogue summarization is a long-standing task in the NLP field and has recently gained traction through the emergence of novel data sets (Gliwa et al., 2019; Zhu et al., 2021) and community efforts like the AutoMin¹ shared task or the SummDial@SIGDial 2021 special session². Dialogues can take a wide variety of forms ranging from formal interviews on a specific topic, political debates to informal conversations over the telephone.³ Therefore, the question of what is a suitable, appropriate summary of this type of data emerges. While abstractive and extractive summaries have emerged as the de-facto standard for summarization of continuous text (either single or multiple documents), the situation is less clear for dialogue summarization: What is a “proper” summary for the different types of dialogues?

Several dialogue corpora with associated human-written summaries have been created. These summaries differ significantly in type, style, and focus, depending on the instructions that were given to the human annotators. It is usually not clear why the particular summary type was chosen for the dialogue corpus at hand. In fact, we are not aware of any well-founded theory that answers this questions.

To close this gap, we leverage the well-established linguistic model of dialogue types by Walton and Krabbe (1995) to identify suitable summary items for the different types of dialogues. This results in a combination of linguistically defined dialogue types, their features, and the suitable summary items. We then place all existing dialogue data sets with summaries that we are aware of into this matrix. This allows us to map the available resources and to identify gaps, which opens up directions for future work.

More precisely, this work presents four contributions:

1. A concise presentation of the linguistically grounded classification of dialogue types by Walton and Krabbe (Section 2)
2. A mapping from dialogue types to potential summary items and associated NLP tasks (Table 2). This indicates which summaries would be appropriate for which dialogue type.
3. An overview of all existing data sets for dialogue summarization that we are aware of (Section 3), which will be useful for researchers in the field even independent from the linguistic model.
4. A mapping from the existing data sets to the linguistic model, and an analysis of potential resource gaps (Section 4).

We also present the overview of existing dialogue summarization data sets in a comprehensive tabular overview in Table A in the Appendix.

¹https://elitr.github.io/automatic-minuting/index.html
²https://elitr.github.io/automatic-minuting/summdial.html
³In this paper we focus on spontaneous spoken dialogues, leaving out written dialogues such as Twitter discussion, scripted dialogues, which occur for example in movies, and the summarization of material spoken by single persons only.
2 Dialogue Types in Linguistics

The analysis of dialogues within linguistics is mainly investigated in the fields of conversation analysis and pragmatics. A large body of work investigates speech acts (Searle, 1969; Grice, 1975, inter alia), i.e., dialogues are decomposed into individual turns and their communicative intents are analysed. A smaller body of work focuses on establishing a typology of dialogues. Among these works, Walton and Krabbe (1995) is a well-established model that is often cited and discussed.

2.1 The Walton & Krabbe Model

Walton and Krabbe developed a model of dialogues types and their features which is often picked up in subsequent work in various fields. Table 1 (upper part) shows the model. It features six basic dialogue types: Persuasion, Negotiation, Inquiry, Deliberation, Information-seeking, and Eristics. There are three additional mixed types: Debate (Persuasion and Eristics), Committee meeting (mainly Deliberation), and Socratic Dialogue (mainly Persuasion).4

Recently, Macagno and Bigi (2018) showed how Walton and Krabbe’s model is connected to theories of speech acts, dialect acts, and pragmatic acts and concepts such as "communicative intentions". Walton and Krabbe’s dialogue types were explored by research in multi-agent communication in computer science. For example, Reed (1998) applies the model to derive dialogue frames to describe multi-agent interactions.

A related approach to dialogue type categorization is presented in Franke (2010, 2011). The approach develops a taxonomy of (minimal) dialogues. Minimal dialogues are sequences of speech acts in a dialogue that have ended in a conclusion or decision. A naturally-occurring dialogue is then modelled as a sequence of these minimal dialogues.

It is noteworthy that naturally-occurring dialogues are seen as a mixture of multiple dialogue types in both aforementioned models. Still, we find that most of the dialogue corpora we examine in Section 3 can be assigned to one (or two) main dialogue type(s) under the Walton and Krabbe model.

We choose the Walton and Krabbe model as the basis of our analysis of resources in the dialogue summarization space as it is generally the most established one and has been shown to extend well into other domains. Furthermore, the features that Walton and Krabbe attribute to the dialogue types enable us to infer desiderata for the type of summary that suits the dialogue type. For example, if the main goal of a negotiation is "making a deal", then a suitable summary would present the deal that resulted from the negotiation. Similarly, if the main goal of a debate is "accommodating conflicting points of views", then a suitable summary would list these points of view, and by extension, attribute them to the speakers participating in the debate, and, going further, provide insight into the reasoning of the speakers etc. Finally, the generation of the desirable summary types can then be decomposed naturally into well-established NLP tasks such as topic detection, argument mining, and stance detection, etc.

In summary, the Walton and Krabbe model and its features provide a structured perspective on dialogues that lets us identify suitable connections between dialogue types and summary items, and enables us to pin-point NLP tasks that are applicable for accomplishing such summaries.

2.2 Mapping Dialogue Types to Summary Items

Having selected the model of dialogue types by Walton and Krabbe (1995) as the lens through which we wish to explore the resources in the dialogue summarization domain, we first infer desirable properties of summaries for each of the dialogue types. For this purpose, we examine the dialogue types’ features (primarily: Initial situation and Main goal; secondarily: Participant’s aim and Side benefits) to derive items that an optimal summary would contain in this view. To link the desirable summary items to specific NLP tasks, we note down NLP targets that need to be identified and extracted to enable a summarization system to produce the summary items in its outputs.

The lower part of Table 1 presents the result of this process.5 The summary items are ordered by importance in relation to our prioritization of the dialogue type features (i.e. Main goals are more important than Side benefits). We exemplify our mapping based on the Persuasion dialogue type:

---

4We omit the mixed dialogue types in Table 1 for brevity, as they are combinations of the other types.

5To encourage different takes in this mapping process, the authors of this paper individually performed the task of mapping dialogue types to summary items and NLP tasks and then held a discussion to harmonize the mappings. Overall, the mappings of the authors overlapped to a large extent and complemented each other, i.e., no conflicting points or disagreement emerged.
Walton and Krabbe’s model is connected to theo-

Walton and Krabbe developed a model of dialogues (mainly Persuasion) that most of the dialogue corpora we examine in Section 3 can be assigned to one (or two) main dia-

logues. Minimal dialogues are sequences of speech acts in a dialogue that have ended in a conclusion. There are three additional mixed types: De-

er-ers/Losers, Writers/Lossers, Controversies, positions, Arguments, Win-

ners/Losers, Controversies, Final deal, Initial inter-

ests, Winners/Losers, Evolution of deal, Argumenta-

ions, Initial inquiry, Gained/new knowledge, Reached agree-

ment, (Line of) arguments, Mentioned facts, Decision, Initial need for ac-

tion, Positions of speakers, Evolution of decision, Win-

ners/Losers, Emotions, Initial problem, Solution, Posi-

tions, Emotions, Initial conflict, Resolution/agreement, Win-

ners/Losers, Arguments, Emotions, NLP targets

| Summary items | POVs, Resolutions, Disagreements, Positions, Arguments, Winners/Losers, Controversies | Final deal, Initial interests, Winners/Losers, Evolution of deal, Arguments, Initial inquiry, Gained/new knowledge, Reached agreement, (Line of) arguments, Mentioned facts, Decision, Initial need for action, Positions of speakers, Evolution of decision, Winners/Losers, Emotions, Initial problem, Solution, Positions, Emotions, Initial conflict, Resolution/agreement, Winners/Losers, Arguments, Emotions |

Table 1: Categorization of dialogue types (columns) and their features (rows) according to Walton and Krabbe (1995), and their mapping to our proposed summary items (sorted by importance) and the applicable NLP tasks’ target information.

The main goal of Persuasion dialogues is to resolve a conflict between multiple speakers. Each particip-

ant wants to persuade the others. For a summary, we are mainly interested in the different conflicting points of view (POV) and the resolution of the disagreement. However, the arguments used to resolve the conflict, and the final “winner” are also of interest. For each of these summary items, a corresponding NLP task can be used to extract a specific item. For instance, to extract the different POVs, stance detection can be applied. To extract the arguments used to persuade others, argument detection is applicable, etc. That is, summaries of a dialogue under a given dialogue type would ideally include these targets explicitly in a structured manner to facilitate the creation and evaluation of automatic summarization systems.

The list of all NLP targets emerging in the mapping are: Topics (tracking), Decisions/Action items, Arguments, Emotions/Sentiment, Stances, Keyfacts, and Knowledge. We will apply this inventory of NLP targets in Section 4 to map out existing re-

sources and investigate which summary items have been explored for which dialogue types.

3 Data Sets – An Overview

We next provide an overview of existing dialogue summarization data sets. The overview is comple-

mented by Table A in the Appendix which offers a compact and comprehensive outline of the data sets including descriptions, sizes, covered languages, and available summary types. We divide the data sets into the domains that they cover (Meetings, Broadcast Interviews, Customer and Patient Support, Spontaneous Conversation) and discuss applicable dialogue types.

Dialogues can be either spoken or written. While several corpora of written or more formal dialogues
and their summarization have emerged recently (Gliwa et al., 2019; Chen et al., 2021, inter alia), we here focus on corpora for summarization of (transcripts of) spoken dialogues, which is considerably different than summarization of text, as described for example in Gurevych and Strube (2004).

Work on summarizing spoken dialogues (i.e. involving more than one speaker) started in the late 1990s and early 2000s (see for example Zechner and Waibel (2000b,a)). These already covered a great variety of different types of dialogues, such as TV discussions (NewsHour, CNN CrossFire), phone calls (CALLHOME, CALLFRIEND) and meetings. An overview of these early approaches into summarizing dialogues can be found in Zechner (2002).

At the same time, the VERBMOBIL project, which focused on negotiations dialogues, also worked on summarising these (Reithinger et al., 2000; Alexandersson et al., 2000).6

3.1 Meetings

The topic of summarizing meetings gained considerable attraction with extensive work on the ICSI-Corpus (Morgan et al., 2001) and the AMI-Corpus (Murray et al., 2007, e.g.). Murray et al. (2005) presented work on manually summarizing the ICSI meetings, where annotators were instructed to "construct a textual summary [...] aimed at someone who is interested in the research being carried out". Four headlines or questions served as guidelines: 1) Why are they meeting and what do they talk about? 2) Decisions made by the group, 3) progress and achievements and 4) problems described. Liu and Liu (2008) extended this work by creating more human summaries and evaluating the summaries based on a questionnaire to be filled out by humans. Other work looked in more detail into how to detect and summarize action items, their descriptions and their appropriate time frames (Purver et al., 2007, e.g.).

The AMI corpus was also extensively studied in the context of summarization. However, while the ICSI corpus contains actual meetings of the participating research groups, which had a varied number of participants, the AMI corpus contains meetings of four persons with different roles in a product design scenario, which was not a natural scenario for the participants. Additionally, the topic is always the same, whereas the ICSI corpus has a wide variety of topics that were discussed in the meetings, including for example chit-chat among team members waiting for everyone to arrive. Summaries for the AMI corpus were created in an abstractive way, based on dialogue acts supporting the information in the summaries (Murray et al., 2007).

Fernández et al. (2008) aimed at identifying "decision-making sub-dialogues" in the AMI meeting data. The authors state that a decision sub-dialogue consists of three components: a) an issue raised, b) proposals are considered and c) the decision. To that end, they annotate dialogue acts in the data that represent either the issue, or parts of the resolution and the decision.

Similar to the development in the text summarization domain, the dialogue summarization domain moved to using queries to represent the information need of a specific user (Mehdad et al., 2014). Unfortunately, there was not data created for this scenario and the qualitative evaluation was performed on a small subset of the data.

Wang and Cardie (2012) and Wang and Cardie (2013) also work on summarizing meetings, but rather than aiming for a generic summary, they present work on summarizing focused summaries, that are based on specific aspects of a meeting, such as decisions, action items etc.

Following in the footsteps of the AMI corpus Yamamura et al. (2016) present a similar dataset for the Japanese language named "Kyutech Corpus", which also includes reference summaries created in the same fashion as the reference summaries for the AMI corpus.

More recently, Zhong et al. (2021) used queries to represent information need when accessing the ICSI and AMI corpora.

Another type of meeting dialogues occur in the political domain. Political debates from the UK’s House of Commons have been used by Vilares and He (2017). The authors aim to produce summaries which give a brief overview on the main viewpoints exchanged and perspectives expressed, which puts it in the area of stance classification and argument mining.

Committee meetings from the Welsh and Canadian Parliament are used by Zhong et al. (2021). Their aim is to create informative summaries based on two types of queries: General queries and specific queries, which included discussion points,
opinions, ideas etc. In the discussions elements relevant to the queries have been annotated, as well as informative summaries created.

**Dialogue Types**  The discussed corpora in the meeting domain mainly cover project, team, and committee meetings. Given the Initial situation settings of *Need for action, conflict of interest & need for cooperation*, and the Main goals *Reach a decision, Making a deal*, we assign this domain to the dialogue types *Deliberation* and *Negotiation*.

### 3.2 Broadcast Interviews

TV discussions were already studied in the early phases of speech summarization. More recent work is presented by Zhu et al. (2021) based on NPR and CNN interviews. Reference summaries are based on the descriptions of the interviews and the list of topics discussed.

Podcasts are another form of exchange, that can be an interview, but it can also be a discussion. Clifton et al. (2020) present a data set of Podcasts used for summarization. Reference summaries are based on creator-generated descriptions, which are most likely rather indicative than informative. Using generic summarization algorithms, summaries are created automatically and evaluated manually.

**Dialogue Types**  While the formats covered in the corpora in this domain are rather open by nature, we map it to the dialogue types *Information-seeking*, e.g., interviews with an experts where ignorance (Initial situation) is remedied by the expert's knowledge (Main goal), and *Debate*, where the Initial situation is the presence of conflicting views that are accommodated and discussed in front of an audience (Main goal).

### 3.3 Customer and Patient Support

Early work in dialogue summarization also includes call-center dialogues. Higashinaka et al. (2010) present work in this direction, which is unfortunately not based on actual call-center dialogues, but rather on recordings of people who were assigned various roles. Tamura et al. (2011) improved on this by using actual call center data. As the logs available for each dialogue were deemed unsuitable for summarization, two types of summaries were created: 1) Indicative summaries, for agents or managers to grasp the gist of the calls and 2) Informative summaries, that contain the content and allow managers to get necessary details of the calls.

Favre et al. (2015) also present work on summarizing call center dialogues. The aim is to create synopses of the calls, which contain the problem and the suggested solution. As opposed to most other work presented, the data set covered not only English, but French (Decoda Corpus) and Italian (Luna Corpus). Based on the same data sets Danieli et al. (2016) looks into analysing the behavior shown in the conversation, which is an important aspect for quality assurance supervisors.

Liu et al. (2019) present work on the DiDi corpus, which contains dialogues from customer service centers and summaries created by the respective agents. Their aim is to identify key-point sequences in the dialogues, to which end they devise a tagging system with 51 labels, ranging from "Question Description" to "Solution".

Zhao et al. (2020) present work on the Automobile Master Corpus, which contains data from a customer question and answer scenario. It is unclear what the summaries are aimed at, so we have to assume that they are generic summaries.

Various data sets have been used for summarization that come from the medical domain. Acharya et al. (2019) present work on a data set where patients with a specific condition are interviewed. As the data contains actual interviews it cannot be shared. The summaries created aim to include sentences that motivate patients to get better.

Joshi et al. (2020) and Yim and Yetisgen (2021) work on a data set of medical interviews where reference summaries are created by medical doctors, instructing them to summarize as they would for a "clinical note by including all the relevant information". A specific focus was put on negative utterances such as "does not have symptom X".

**Dialogue Types**  This domain clearly evolves around the need for specific information exchange (Initial setting) and passing knowledge between the speakers (Main goal). We thus assign it the *Information-seeking* dialogue type.

### 3.4 Spontaneous Conversations

Spontaneous or rather informal conversations were already part of the early work presented by Zechner and Waibel (2000b) and Zechner and Waibel (2000a), which looked at the CALLHOME and CALLFRIEND data, which consists of telephone conversations.

A similar setting is the basis for the Switchboard Corpus, which also contains telephone con-
versations on specific topics. Gurevych and Strube (2004) required annotators to "select the most important utterances" in a selection of dialogues and formed two types of gold standard: One based on all three annotators and one based on annotations by at least two annotators.

A more recent type of informal dialogues has been presented by Rameshkumar and Bailey (2020) which contains dialogues in the context of pen and paper role-playing games (CD3 data set). Summaries are provided through a wiki and are produced by fans of the associated show.

Dialogue Types This domain is difficult to assert in terms of dialogue types as the features Initial situation and Main goal are not clearly identifiable. While speakers were given a specific topic for a conversation in most cases, they were not specifically instructed to converse in a predefined manner. We can hence only speculate on the dialogue types mirrored in these conversations; the conversations would have to be examined individually to determine a sequence of matching dialogue types, which is infeasible in our study.

4 Mapping Data Sets to Summary Items

Given the overview of dialogue summarization resources and their mapping to dialogue types under the linguistic model in the previous section, and the summary items assigned to the dialogue types in Section 2.2, we are now able to tabulate the corpora and the summary items to see what areas in this space are covered and where there are opportunities for future work.

Specifically, we tabulate corpora and the NLP targets that are mapped to the summary items and insert the paper references that cover the summary item for a given corpus. We perform this mapping under the requirement that a resource explicitly annotates a given NLP target in a structured manner. That is, while a general, abstractive, manual summary of a meeting might include e.g. action items or decisions, they might not be marked explicitly as such in the summaries or the underlying transcripts. In such a setting, the resource would not enable the creation of summarization systems that explicitly extract e.g. action items.

Table 2 shows the result of this mapping. A quick glance reveals that only a small portion of the potential NLP targets are explicitly annotated in the summarization resources. The table also shows where efforts to create resources have been focused in the dialogue summarization space: The corpora in all domains mainly offer topics-related summaries. The meetings domain is an exception, where considerable effort has been put into annotating decisions and action items.

5 Discussion

Early approaches to create resources for dialogue summarization in the 2000’s were based on spontaneous conversations. Such dialogues are difficult to map to the Walton and Krabbe types, as the features instantiations, such as Initial Situation or Main Goal, are hard to determine. The diversity of these conversations also makes it difficult to define clear guidelines for creating summaries: Annotators were mostly guided by a somewhat under-specified relevancy criterion and were given a length constraint. In regards to the covered summary items, such extractive summaries might contain e.g. decisions and stances etc., however, they are not marked or labeled in the extracted dialogue segments explicitly.

In the Meetings domain, summarization efforts became more specific and a substantial body of work looked into decisions and action items, which resulted in structured data sets for these summary items. For other summary items that the dialogue types Negotiation and Deliberation yield, such as Stances and Arguments, no structured resources exist, however.

Available summaries in the Broadcast domain consist of content description by the authors/creators of the content, i.e. they were not created by researchers for the purpose of dialogue summarization. The descriptions thus rather follow the (potentially commercially-motivated) goal of raising interest in a audience, rather than providing an informative or indicative summary. The communicative intent of such descriptions can therefore be considered to be substantially different from that of research-oriented summarization data sets. Naturally, such content descriptions do not explicitly make available any specific summary items.

We omit the Knowledge summary item, since no resource covers it. However, Knowledge Discovery might be an interesting task in Inquiry dialogues.

However, it is not always straight-forward to apply the NLP targets to resources. For example, in the QMSum corpus, the most important topics are summarized for all dialogues, but the queries that cover decisions are not guaranteed to be present for all dialogues and are not explicitly labeled as being related to decisions.
In the customer and patient support domain, summarization efforts also leveraged readily available resources such as synopses of call logs or doctor’s notes as the summarization targets. Here, the goal of summarization efforts can be mainly described as automating the task of manually producing such notes or synopses. Hence, many linguistically motivated summary items that our approach yields for the Information-seeking dialogue type may simply not apply to the particular use cases that are covered by the existing resources, and are thus not marked explicitly as such.

6 Conclusion

We have provided an overview of existing corpora in the domain of spoken dialogue summarization. We found that topic-related extractive or abstractive summaries are predominant, and are often guided...
by high-level criteria, i.e. summary guidelines ask for content of "high relevancy" to be included without further specifications.

Furthermore, we have applied a linguistically motivated view on dialogues to the available corpora that yields more specific summary items, such as arguments, stances, or emotions. We found that such specific items are scarcely available in a structured manner in existing corpora. As there are several resources available for e.g. argument mining (Lawrence and Reed, 2020) and stance detection (Küçük and Can, 2020) in dialogues, a potential direction for future work could be an effort to bring together such resources.

While our model-driven view on the dialogue summarization space might be insightful and fruitful for future research, it should not be understood in a normative way: it is not intended to point out that certain directions are misguided. For instance, although our mapping does not yield Emotion as a summary item for Negotiation dialogues, there might be relevant use cases for this line of inquiry. Neither does the approach have any claim to completeness in terms of meeting the information needs of different users. In this regard, query-based approaches seem to hold a large potential to cover a wide variety of information needs (Zhong et al., 2021). However, since summary items are seamlessly embedded in the natural-language responses in such settings, it is uncertain how well query-based methods are able to generate on-the-fly responses for realistic queries like "what are the action items assigned to me and by when do I have to complete them?". Answering such information needs robustly seems to necessitate that the underlying information is extracted in a structured manner (Purver et al., 2007, e.g.) to be able to generate an appropriate and complete response.

Overall, our analysis indicates that the question of what are appropriate summaries of dialogues is a challenging one, and we have presented a view that offers some answers. While emerging query-based approaches seem to be a fruitful direction due to their potential to cover a high variety of information needs, we believe that linguistic considerations, as those outlined in this work, should also be leveraged to support resource creation efforts in the dialogue summarization space in future work.

Future work should also evaluate the summaries resulting from more dialogue-specific annotations as opposed to generic summaries especially with respect to the individual information needs of various users. This of course also leads to developing methods that take the information need into account when creating such summaries automatically.

References


Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. 2008. Mod-


Klaus Zechner and Alex Waibel. 2000b. Minimizing word error rate in textual summaries of spoken language. In 1st Meeting of the North American Chapter of the Association for Computational Linguistics.


A Appendix
Meetings Corpora. Dialogue types: Negotiation, Deliberation

<table>
<thead>
<tr>
<th>CORPUS</th>
<th>DESCRIPTION</th>
<th>LANG</th>
<th>SUMMARY CONTENTS</th>
</tr>
</thead>
</table>
| VerbMobil       | Negotiations in the domains of scheduling, travel planning, and hotel reservations | DE, EN, JP | - Agreements on locations, dates, hotels, trains (Reithinger et al., 2000).  
- Agreements on scheduling, accommodation, traveling, entertainment. (Alexandresson et al., 2000). |
| ICSI Corpus     | Informal, natural, and even impromptu meetings at ICSI. 38 meetings for a total of 39 hours, transcribed about 12 hours. 237 participants, 49 unique speakers. | EN   | - Summaries answering the following questions: Why are they meeting and what do they talk about? Decisions made by the group? Progress and achievements? Problems described (Murray et al., 2005).  
- Find dialogue acts that relate to action items (descriptions, time frames, owners, agreements) (Purver et al., 2007).  
- Abstract summarizing each important output for every meeting. Decision and problem summaries are annotated (Wang and Cardie, 2013). |
| AMI Corpus      | 100 hours of meeting recordings.                          | EN   | - Ranking the dialogue acts in terms of being extract-worthy (Murray et al., 2007).  
- Classify utterances related to decisions: issue (I), resolution (R), and agreement (A). Two authors annotated 9 and 10 dialogues each (Fernández et al., 2008).  
- An abstract summarizing each decision; dialogue acts that support each decision are annotated (Wang and Cardie, 2012).  
- Abstract summarizing each important output for every meeting. Decision and problem summaries are annotated (Wang and Cardie, 2013). |
| Kyutech Corpus  | A decision-making task in a virtual shopping mall in a virtual city. 9 conversations. | JP   | - Abstractive manual summaries as in the AMI corpus (Yamamura et al., 2016). |
| QMSum           | AMI, ICSI, and 25 committee meetings of the Welsh Parliament and 11 from the Parliament of Canada | EN   | - Select and summarize relevant spans of meetings in response to a query (Zhong et al., 2021). |
| AutoMin         | Technical meetings and parliamentary proceedings.         | EN, CZ | - Meeting minutes (paper in print; https://elitr.github.io/automatic-minuting/index.html) |

Broadcast Corpora. Dialogue types: Information-seeking, Debate

<table>
<thead>
<tr>
<th>CORPUS</th>
<th>DESCRIPTION</th>
<th>LANG</th>
<th>SUMMARY CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MediaSum</td>
<td>Interview transcripts from NPR and CNN. 49.4K NPR transcripts and 414.2K from CNN.</td>
<td>EN</td>
<td>- Topic descriptions as summaries (Zhu et al., 2021).</td>
</tr>
<tr>
<td>Spotify Podcast Dataset</td>
<td>100,000 podcast episodes, comprising ~60,000 hours of speech.</td>
<td>EN</td>
<td>- Creator-generated descriptions as reference summaries (Clifton et al., 2020).</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>CORPUS</th>
<th>DESCRIPTION</th>
<th>LANG</th>
<th>SUMMARY CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiDi</td>
<td>Logs in the DiDi (mobile transportation platform) customer service center.</td>
<td>EN</td>
<td>- Abstractive summaries written by agents. ~300k pairs of dialogues and summaries. &quot;Key point sequences&quot;, i.e. a set of 51 action/decision items are also annotated (Liu et al., 2019).</td>
</tr>
<tr>
<td>Call center I</td>
<td>Simulated contact center dialogues in six domains. 15–20 tasks per domain. ~700 dialogues.</td>
<td>JP</td>
<td>- Scenario texts used as reference data (Higashinaka et al., 2010).</td>
</tr>
<tr>
<td>Call center II</td>
<td>4,596 call logs from a Japanese contact center.</td>
<td>JP</td>
<td>- 1. Indicative Summary: Extract utterances to grasp the gist of calls. 2. Informative Summary: Utterances to grasp the details of calls (Tamura et al., 2011).</td>
</tr>
<tr>
<td>CCCS</td>
<td>Conversations from the Decode and Luna corpora of French and Italian call centre recordings. Recordings duration from a few to 15 minutes. 100 conversations in EN, FR each, translated to EN.</td>
<td>FR, IT, EN</td>
<td>- Abstractive summaries about the main events of the conversations, such as the objective of the caller, whether and how it was solved by the agent, and the attitude of both parties. Synopses written by quality assurance experts from call centres (Favre et al., 2015).</td>
</tr>
<tr>
<td>Data Set</td>
<td>Description</td>
<td>Language(s)</td>
<td>Notes</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------------</td>
<td>-------</td>
</tr>
<tr>
<td><strong>Telemedicine</strong></td>
<td>25,000 conversations from a telemedicine platform.</td>
<td>EN</td>
<td>- Medical doctors were asked to summarize the sections of 3000 snippets as they would for a typical clinical note by including all the relevant information (Joshi et al., 2020).</td>
</tr>
<tr>
<td><strong>Clinical Encounter Visits</strong></td>
<td>Audio and clinical notes from clinical encounter visits from 500 visits and 13 providers.</td>
<td>EN</td>
<td>- Clinical notes as summary of the patient visit (Yim and Yetisgen, 2021).</td>
</tr>
<tr>
<td><strong>Spontaneous Conversation Corpora</strong></td>
<td><strong>Dialogue types:</strong> N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Callhome corpus</td>
<td>Spontaneous telephone conversations.</td>
<td>EN, ES</td>
<td>- For 9 English and 14 Spanish dialogues, the most relevant turns were marked (Zechner and Waibel, 2000b).</td>
</tr>
<tr>
<td>Television shows</td>
<td>Four audio excerpts from four television shows.</td>
<td>EN</td>
<td>- Most relevant, meaningful, concise, and informative phrases (Zechner and Waibel, 2000a).</td>
</tr>
<tr>
<td>Switchboard</td>
<td>Telephone conversations of at least 10 minutes duration on a given topic. ~2000 turns.</td>
<td>EN</td>
<td>- 10% of all utterances in the dialogue marked as being relevant (Gurevych and Strube, 2004).</td>
</tr>
<tr>
<td>DialogSum</td>
<td>Combination of English learner corpora and dialogue understanding data sets. 13,460 dialogues.</td>
<td>EN</td>
<td>- (1) convey the most salient information; (2) be brief (no longer than 20% of the conversation); (3) preserve important named entities within the conversation; (4) be written from an observer perspective; (5) be written in formal language (Chen et al., 2021).</td>
</tr>
<tr>
<td>CRD3</td>
<td>Transcripts of Dungeons and Dragons role-playing game. 398,682 turns.</td>
<td>EN</td>
<td>- Multiple summaries available, e.g. an abstract of the resulting plot/narrative of a game. Includes abstractive summaries collected from the Fandom wiki (Rameshkumar and Bailey, 2020).</td>
</tr>
</tbody>
</table>

Table 3: Overview of existing dialogue summarization data sets. The last column lists papers that provide manually created summaries for a given corpus.
SwissText 2021
Swiss Text Analytics Conference 2021

Proceedings of the Swiss Text Analytics Conference 2021
Winterthur, Switzerland, June 14-16, 2021 (held online due to COVID19 pandemic).

Edited by
Fernando Benites *
Don Tuggener **
Manuela Hürlimann **
Mark Cieliebak **
Manfred Vogel *

* FHNW University of Applied Sciences Northwestern Switzerland, Institute for Data Science, 5210 Windisch, Switzerland
** ZHAW Zurich University of Applied Sciences, Institute of Applied Information Technology (InIT), 8401 Winterthur, Switzerland
Table of Contents

- Preface
  Summary: There were 11 papers submitted for peer-review to the Scientific Track. Out of these, 7 papers were accepted for this volume. Submissions to the Applied Track were also reviewed by the program committee. Submissions to the shared tasks were reviewed by the task organizers.

Scientific Track

- Re-Evaluating GermEval17 Using German Pre-Trained Language Models
  Matthias Aßenmacher, Alessandra Corvonato, Christian Heumann

- Emotions in the parliament: Lexical emotion analysis of parliamentarian speech transcriptions
  Egon Werlen, Christof Imhof, Per Bergamin

- Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus
  Michel Plüss, Lukas Neukom, Christian Scheller, Manfred Vogel

- Predicting COVID-19 Cases using Reddit Posts and other Online Resources (short paper)
  Felix Drinkall, Janet Pierrehumbert

- NFinBERT: A Number-Aware Language Model for Financial Disclosures (short paper)
  Hao-Lun Lin, Jr-Shian Wu, Yu-Shiang Huang, Ming-Feng Tsai, Chuan-Ju Wang

- Neural Argumentation Mining on Essays and Microtexts with Contextualized Word Embeddings (short paper)
  Mohammad Yeghaneh Abkenar, Manfred Stede, Stephan Oepen

- Old is Not Always Gold: Early Identification of Milestone Patents Employing Network Flow Metrics
  Manajit Chakraborty, Fabio Crestani

Shared Task: Sentence End and Punctuation Prediction in NLG Text

- The Sentence End and Punctuation Prediction in NLG Text (SEPP-NLG) Shared Task 2021
  Don Tuggener, Ahmad Aghaebrhimian

- UZH OnPoint at Swisstext-2021: Sentence End and Punctuation Prediction in NLG Text Through Ensembling of Different Transformers (short paper)
  Andrianos Michail, Silvan Wehrli, Terézia Bucková

- Multilingual Simultaneous Sentence End and Punctuation Prediction (short paper)
  Ricardo Rei, Fernando Batista, Nuno M. Guerreiro, Luisa Coheur

- FullStop: Multilingual Deep Models for Punctuation Prediction
  Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, Hans-Joachim Böhme

- University of Regensburg @ SwissText 2021 SEPP-NLG: Adding Sentence Structure to Unpunctuated Text (short paper)
  Gregor Donabauer, Udo Kruschwitz

- Participation of HULAT-UC3M in SEPP-NLG 2021 shared task (short paper)
  Jose Manuel Masiello-Ruiz, Jose Luis Lopez Cuadrado, Paloma Martinez
Deep Contextual Punctuator for NLG Text (short paper)
Vandan Mujadia, Pruthwik Mishra, Dipti Misra Sharma

Shared Task: Swiss German Speech to Standard German Text

SwissText 2021 Task 3: Swiss German Speech to Standard German Text (short paper)
Michel Plüss, Lukas Neukom, Manfred Vogel

Dialectal Speech Recognition and Translation of Swiss German Speech to Standard German Text: Microsoft’s Submission to SwissText 2021 (short paper)
Yuriy Arabskyy, Aashish Agarwal, Subhadeep Dey, Oscar Koller

ZHAW-CAI: Ensemble Method for Swiss German Speech to Standard German Text (short paper)
Malgorzata Anna Ulasik, Manuela Hürlimann, Bogumila Dubel, Yves Kaufmann, Silas Rudolf, Jan Deriu, Katsiaryna Mlynchyk, Hans-Peter Hutter, Mark Cielliebak

2021-09-13: submitted by Don Tuggener, metadata incl. bibliographic data published under Creative Commons CC0
2021-09-27: published on CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073) [valid HTML5]
Welcome to the 6th Swiss Text Analytics Conference (SwissText 2021)

The SwissText conference series was launched by the Zurich University of Applied Sciences in 2016. The first edition was already a huge success with more than 170 participants. In 2020, the conference was handed over to the Swiss Association for Natural Language Processing (SwissNLP), which now organizes the conference in collaboration with different Swiss universities.

This year, SwissText was hosted at the School of Engineering of the University of Applied Sciences and Arts Northwestern Switzerland (FHNW), and co-organized by the Zurich University of Applied Sciences (ZHAW).

Due to the ongoing Corona crisis, we were again forced to hold SwissText as an online conference. Based on our experience from 2020 -- where we had to switch to the online setting at very short notice -- we were better prepared this time and included several activities in the program that allowed the participants to interact and network in the online setting: a “Battle of NLP Ideas”, “Surprise Networking” for NLP experts, and a virtual poster and sponsor exhibition. In addition, there were 34 presentations in the form of talks, posters, and demos.

The conference was accompanied by three shared tasks:

- Sentence End and Punctuation Prediction in NLG Text
- Swiss German Speech to Standard German Text
- Text Normalization for Swiss German

The results of the first two shared tasks were presented at the conference, while the third unfortunately did not receive any submissions. In addition, SwissText hosted three pre-conference workshops:

- 2nd European Language Grid (ELG) Workshop
- NLP in Finance
- NLP efforts against COVID-19 in Switzerland

We received 38 submissions for scientific, applied, and demo presentations. You will find the full scientific papers and abstracts of the applied talks and posters in separate sections in these proceedings. A great addition to the conference program were the “Highlights Talks”, where researchers presented outstanding ideas and insights in new NLP approaches from other top conferences.

We would like to thank our keynote speakers Sebastian Welter (Accenture), Lucia Specia (Imperial College London), Lewis Tunstall (Huggingface), Zenodia Charpy and Adam Grzywaczewski (both NVIDIA). Their perspectives and contributions are much appreciated. A big thank you also to our workshop organizers and the organizers of the shared tasks, and to all members of our programme committees for their excellent work.
We are grateful to our sponsors and partners, who supported us in this unusual setting. In particular, we would like to acknowledge the generous support by Innosuisse (the Swiss Innovation Agency) and our co-organizer, the data innovation alliance.

Last but foremost, we would like to thank Manuela Hürlimann and Fernando Benites, who were the main organizers. They were assisted by Michel, Marc, Don, Claudio, Yanick, Alla, and many more for various tasks. Without them this conference would not have been possible! Further, we are thankful for the support of the programme committees for the scientific and applied track.

It was a great pleasure for us to organize and chair this conference. We hope that all participants enjoyed the conference as much as we did!

Mark Cieliebak and Manfred Vogel
Conference Chairs
Evaluation of Dialogue Systems

Thesis
presented to the Faculty of Arts and Social Sciences of
the University of Zurich
for the degree of Doctor of Philosophy

by
Jan Milan Deriu

Supervisors:
Prof. Dr. Martin Volk, University of Zurich
Prof. Dr. Eneko Agirre, University of the Basque Country
Prof. Dr. Mark Cieliebak, Zurich University of Applied Sciences

Zurich, 2021
Abstract

We investigate evaluation methods for dialogue systems. We focus on conversational dialogue systems and question-answering dialogue systems since each class poses different challenges. For conversational dialogue systems, we tackle the inefficiency and unreliability of the evaluation process. For question-answering dialogue systems, we tackle the problem of evaluation after deployment. The main contributions are:

- We introduce a novel paradigm for evaluating conversational agents based on the bot-to-bot talk. This new paradigm allows sampling conversations automatically, reducing human involvement in the evaluation process. We apply this paradigm to two scenarios - first, Spot The Bot, an evaluation procedure based on bot-bot talk. Humans read automatically generated dialogues and decide for each interlocutor, whether it is a human or a bot. Based on this feedback, the bots are ranked. We show that the rankings are robust and reproducible. The second scenario is AutoJudge, an automated metric for evaluating conversational dialogue systems. It is trained on automatically generated dialogues that are annotated on the turn level by human judges. We show that AutoJudge achieves good correlation scores with humans and that it can be used as a meta-selection model to select the best answer from different dialogue systems.

- We introduce a novel evaluation procedure for question-answering dialogue systems over databases, which allows evaluation after deployment. That is, the system can be evaluated without the need for a gold-standard reference. The method is based on back-translating the generated SQL query to a synthetic question. Textual semantic similarity is applied to the original user input and the synthetic question to determine if the underlying SQL query is correct.

- We introduce a novel annotation procedure to generate pairs of questions and queries more efficiently. The procedure is based on inverting the process, i.e., we first sample a structured query from a context-free grammar, and then humans write the corresponding question. We show a 4-fold improvement in the time needed to generate data with respect to traditional approaches where experts write both the question and the query.
1 Introduction

This thesis treats the task of evaluating dialogue systems, a crucial step during dialogue system development, which is still unsolved. The underlying question that we ask when evaluating a dialogue system is if the dialogue was of sufficient quality. Different definitions of dialogue quality depend on the context and scope of the dialogue system under consideration. We define three types of dialogue systems in research: task-oriented systems, conversational dialogue systems (also known as chit-chat bots), and question-answering dialogue systems [Deriu et al., 2020b].

- Task-oriented dialogue systems are developed to solve a task. For instance, book a flight from Madrid to London on a given date. They are characterized by a highly structured dialogue, which does not allow for dialogues out of scope.

- Conversational dialogue systems (also called chit-chat bots or social bots) are developed to engage the user in a conversation. There is no strictly defined goal. They are characterized by conversations that allow for more variety in different topics, such as talking about the weather or trivia.

- Question-Answering (QA) dialogue systems are developed to answer questions by the user. There are different types of question-answering systems. For instance, extractive QA systems that allow for questions about any topic [Choi et al., 2018; Campos et al., 2020] (e.g. Google¹). Natural Language Interfaces to Databases (NLIDB) that answer questions on the contents of a structured database [Affolter et al., 2019].

For completeness sake, we note that this thesis focuses on dialogue systems developed in a research context. In an industrial setting, the different dialogue system types are more integrated than in research settings. For instance, Siri² acts as a conversational dialogue system and a question-answering system. Another example is XiaoIce [Zhou et al., 2020] that allows for all three types of dialogue system behavior depending on the users’ intent. However, these systems are out of scope for this thesis. Here we focus solely on dialogue systems that are developed in the research context. Furthermore, we limit our discussions to dialogue systems that work with written English texts. We exclude dialogue systems that work with different modalities (e.g., speech or gestures), and we exclude dialogue systems trained in languages other than English. We limit this thesis’ scope to dialogue systems that are developed for research purposes for the English language. Further restrictions are mentioned at a more appropriate time.

Depending on the types of dialogue systems, different approaches to evaluations are needed. This is because they all serve different purposes.

¹ https://www.google.ch/
Task-Oriented Dialogue System For instance, a task-oriented dialogue system is deemed high quality if it reaches its goal (e.g., selling a ticket to the customer) efficiently (i.e., with as few interactions as possible) [Schatzmann et al., 2007]. The evaluation of task-oriented systems is automatable as the aforementioned goals can be measured. Usually, a user simulation is developed to automate the evaluation process. One challenge is to model the user simulation as realistically as possible [Deriu et al., 2020b].

Question-Answering Dialogue System Question-answering dialogue systems are of high quality if they can answer many questions correctly. For instance, question-answering systems, which retrieve a span of text from a collection of texts to answer a question, are evaluated by their ability to retrieve the correct span. This is easily measured using F1 scores [Choi et al., 2018; Reddy et al., 2018; Campos et al., 2020]. If the question-answering system is tasked to answer questions over a structured database, the success can be measured using the correctness of the result set or by analyzing the query produced by the system [Yu et al., 2018b;Deriu et al., 2020a].

Conversational Dialogue System On the other hand, conversational dialogue systems are hard to evaluate [Liu et al., 2016;Deriu et al., 2020b]. The fundamental problem lies in what defines a high-quality dialogue and how to operationalize the definition. That is, what is the metric and what is the method to measure this metric? For task-oriented systems, the metric is efficiency and measured by the number of turns needed to reach a goal. For conversational dialogue systems, there exist various metrics that stem from different viewpoints. However, there is not yet an evaluation method, which can measure the metrics reliably and efficiently. The oldest viewpoint is to measure the intelligence of a dialogue system (or artificial intelligence), a view proposed by Turing measured by the Turing Test [Turing, 1950]. Here, a human judge is tasked to talk to a system, which is either a dialogue system or a human that operates a chat interface. When a dialogue system can convince the judge that it is a human, the dialogue system passes the test. A different viewpoint stems from the industry that aims to maximize user engagement. Thus, the dialogue system is evaluated regarding its ability to engage the user in a long conversation. Although the metric (i.e., engagingness) is clear, the method is not straightforward. One possibility is to measure the length of the conversations [Zhou et al., 2020] following the intuition that a long conversation means that the dialogue system is engaging. A different method is to ask humans to interact with the dialogue system and rate if they found the conversation to be engaging [Dinan et al., 2020]. However, it is not clear for each method if it measures the engagingness of the conversation or something different. Another viewpoint stems from the linguistic perspective. Most notably, Grice [1975] introduced four maxims of conversation:

- **Maxims of Quantity (Informativeness):** Ones’ contribution should be as informative as required, not more and not less.

- **Maxims of Quality (Truthfulness):** Ones’ contribution should be true. That means one should not say what he believes to be false or that for which there is no evidence.
• Maxims of Relation (Relevancy): Ones’ contribution should be relevant to the current context of the conversation.

• Maxims of Manner (Clearness): Ones’ contribution should be easy to understand, unambiguous, brief, and orderly.

These maxims were introduced for general conversations and not specifically for dialogue systems. However, a good dialogue system should adhere to these maxims. Current conversational dialogue systems struggle with all these maxims. For instance, they are not truthful, as they tend to change personality from one utterance to the next [Zhang et al., 2018]. Although these maxims serve as a good guideline for conducting a good conversation, they are not easy to measure. For instance, letting humans judge if a contribution is relevant to the context might yield different results depending on the judge. This leads the evaluations to suffer from low inter-annotator agreement making the evaluation unreliable.

**Example** To better showcase the problems of evaluation for conversational dialogue systems, we consider Table 1. This is a conversation with two interlocutors, P1 and P2.

<table>
<thead>
<tr>
<th>Interlocutor</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Why are you so obsessed with online shopping?</td>
</tr>
<tr>
<td>P2</td>
<td>I just do not like crowds of stores.</td>
</tr>
<tr>
<td>P1</td>
<td>You can choose to go to stores where people do not crowd a lot!</td>
</tr>
<tr>
<td>P2</td>
<td>Why are you so concerned with what I like?</td>
</tr>
<tr>
<td>P1</td>
<td>You do not have to restrict your life always on one side.</td>
</tr>
<tr>
<td>P2</td>
<td>Not shopping in stores is not restrictive. I can go ice skating all winter long!</td>
</tr>
<tr>
<td>P1</td>
<td>You have to learn to bear with life situations for the time being.</td>
</tr>
<tr>
<td>P2</td>
<td>I disagree. My parents practice law and I have dealt with enough.</td>
</tr>
</tbody>
</table>

Table 1: Example of a conversation between P1 and P2.

P2. Here, P1 asks P2 about their shopping preferences, and they talk about it. How do we rate this conversation? A critical piece of information is the context. If this dialogue is between two humans, then we rate it differently than when it is between a human and a chatbot since, at the time of writing, we expect a dialogue system to perform worse than a human. If we rate this dialogue with regard to its humanlikeness, one might be tempted to state that P1 and P2 could be humans. However, it is not clear what criteria are used to assess this. On the other hand, if we

---
3 Here, we note that we want to avoid the philosophical discussion of what it means to know the truth and if a dialogue system can know what is true or not.
ask about the level of engagingness of P1, then we might rate P1 as not very engaging since P2 seems to be more and more annoyed by P1s’ behavior. If we apply Grices’ maxims, we can perform a different analysis. We might rate P2s’ level of informativeness as high but not perfect, as in the second to last exchange, P2 starts to talk about ice skating (maybe to divert from the uncomfortable conversation). The same could be said with respect to relevancy. We might rate P1s’ informativeness as low as P1 delivers more information than is required by P2. However, we note that these are the author’s judgments, and the reader might conclude differently.

**Inefficiency.** Another aspect is that the evaluation is inefficient. In most cases, human effort must be leveraged. For instance, humans must interact with the dialogue system, which is a costly and time-consuming task. It is hard to let humans interact with dialogue systems without any instructions and training [Dinan et al., 2020]. Thus, human evaluation needs to be carefully planned, and the judges must be trained to interact with the system. This increases the cost of the evaluation.

A final aspect that we discuss in this thesis is about the time of evaluation. Ideally, a dialogue system would leverage the interaction with humans to improve its capabilities. For instance, in a task-oriented dialogue system, the online feedback can be leveraged to improve a system after deployment [Gašić et al., 2011]. Here, the user states if the task has been successfully solved, which is then translated to a reward function. However, this approach is not easily applicable for the other two types of dialogue systems. For instance, assume a question-answering system for structured data. If the user asks for all the customers who ordered a movie before a given date, the system returns a list of customers. However, the user cannot be sure if the system performed the task correctly since the system might have inserted a wrong date or the wrong movie title.

This Thesis is concerned with two main questions: first, how can the cost of evaluation be reduced, and second, how can a dialogue system leverage live user feedback to self-assess its performance? We restrict the scope to handle neural conversational dialogue systems developed in the research context and question-answering systems that work over structured data, such as SQL databases. The latter systems are also referred to as Natural Language Interfaces to Databases (NLIDB).

### 1.1 Problem Statement for this Thesis

#### 1.1.1 Increased Efficiency of Evaluation

The main issue in evaluating conversational dialogue systems lies in their unreliability and high cost and time intensity. The long-term goal is to develop an automated method that can reliably evaluate a dialogue system. For task-oriented dialogue systems, there already exist methods for evaluation, which reduce human involvement. Question-answering dialogue systems are also evaluated based on automated methods, which measure the response’s correctness to a given question. However, for conversational dialogue systems developed for engaging conversations with humans, there are no reliable methods for evaluation. In fact, automated methods...
(usually based on BLEU score) are shown not to correlate with human judgments at all [Liu et al., 2016]. Even human-based evaluations tend to be unreliable because the aforementioned metrics (e.g., Grices’ maxims) are hard to translate into a reliable evaluation method. This leads human-based evaluations to be prone to low agreement scores and thus, resulting in unreliable comparisons. Thus, the first problem statement in this Thesis is: how can the costs of evaluating a conversational dialogue system be reduced while ensuring that the evaluation is robust and reproducible?

1.1.2 Unreferenced Evaluation after Deployment

The second problem statement is concerned with the evaluation after deployment, where there is no access to labeled data. In machine learning applications, evaluation is often tailored towards the development phase of the algorithm. However, during the deployment phase, the measurement of the live performance is of great interest. Assessing the deployed model’s performance would lead to the ability of live monitoring and opens up to the possibility of automated improvement over time. The main issue is the lack of a gold standard, which can be used for comparison. Thus, the need for so-called unreferenced metrics arises, that is, metrics that do not need a gold reference. Since dialogue systems are designed to interact with humans, the interactive nature of the conversation can be leveraged to automatically get feedback, which can be used to improve the system. Especially for Natural Language Interfaces for Databases (NLIDB), unreferenced methods are of great value since they remove the need to employ costly SQL experts for data annotation. Thus, the second problem statement in this Thesis is: how can NLIDB systems be evaluated without the need for a gold standard, and can this method be used to improve the system automatically?

1.2 Thesis Contributions

The contributions of this Thesis focus on conversational dialogue systems and NLIDBs. This Thesis adds the following contributions:

Bot-Bot Talk as basis for evaluation. We present a novel paradigm for evaluating conversational dialogue systems based on automatically sampled dialogues between two dialogue systems (this also includes self-talk, i.e. when a bot talks to itself). We present two applications, which are based on bot-bot talks. First, Spot The Bot is a framework for human evaluation, where humans are shown dialogues between dialogue systems and need to decide for each interlocutor if it is a human or a bot. The second application is AutoJudge, a trained metric, i.e., a metric trained on examples of rated dialogue turns. These judgments are then used to train a regression model, which learns to judge dialogue turns automatically. The bot-bot talk paradigm allows for evaluating the multi-turn behavior of dialogue systems without generating dialogues between humans and dialogue systems. Thus, both proposals improve efficiency.

Comparative Evaluation of Dialogue Systems. With Spot The Bot, we introduce a novel human evaluation protocol for evaluating conversational dialogue systems efficiently and reliably. The evaluation is based on conversations between bots allowing for a comparative evaluation, which is shown to be more robust. Given a pool of bots,
we randomly sample a set of conversations for each pair of bots through bot-bot talk and let humans find out for each interlocutor if it is a bot or a human. We can then create a ranking for the bots in the pool, which we show to be robust. Since current conversational dialogue systems are not yet human-like enough, we add a time component, which allows us to compare which dialogue system can maintain a human-like appearance for the most amounts of turns.

**Novel Data Collection Methodology for NLIDB.** We introduce a novel data gathering procedure for NLIDB, which increases the efficiency and the amount of content covered in a database. Traditionally, the annotation process is performed manually by letting SQL experts produce pairs of questions and SQL queries or automatically generate question and query pairs based on manually generated templates. We propose to invert the annotation procedure by sampling the queries from a context-free grammar and letting humans write the corresponding question. Our approach produces highly complex questions, which most semi-automated approaches for data generation could not do. To this end, we introduce a novel representation for the SQL queries based on the logical execution plan found in relational databases. It reflects the idea to represent a query as a sequence of operations, where the result of one operation is used as input for the next operation.

**Evaluation after Deployment for NLIDB systems.** We propose an unreferenced evaluation method (i.e., an evaluation method that does not rely on manually annotated data) for NLIDB based on back-translation and textual semantic similarity. To this end, we introduce a rule-based back-translation procedure to represent a query in natural language. We apply this to back-translate the generated query of our NLIDB system. Given the original question and the back-translated question, we apply a textual semantic similarity. This procedure is used as a proxy for evaluating if the NLIDB system correctly answers a question. We show that it can be used as a proxy to evaluate the system.

**Improvement of the NLIDB system.** We propose two applications of the back-translation and semantic similarity evaluation to improve the system’s output based on these results. The first application automatically enhances the training set with automatically generated pairs of questions and queries and uses this synthetic dataset for pretraining. The second application is to rerank the hypothesis of the beam search by semantic similarity scores. We show that the combination of these two applications yields significant improvements.

**1.3 Outline**

First, we review the background of research on the evaluation of dialogue systems in chapter 2. Since the evaluation is tightly coupled with the dialogue systems themselves, we also introduce research on dialogue systems and the relevant corpora for the thesis. Chapter 3 introduces the novel paradigm, namely, bot-talk, as the basis for evaluation. We show its application on both human evaluation and automated evaluation. Chapter 4 introduces the more efficient data collection procedure, the unreferenced evaluation protocol, and the improvement strategy. In chapter 5, we conclude the thesis and present an outlook on future research.
Appendix

A.1 List of Publications

Computer Vision, Perception and Cognition Group


Natural Language Processing Group


## A.2 CAI Team

<table>
<thead>
<tr>
<th>Name</th>
<th>Function</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waqar Ali</td>
<td>Joint doctoral student with University of Venice</td>
<td><a href="mailto:xalw@zhaw.ch">xalw@zhaw.ch</a></td>
</tr>
<tr>
<td>Mohammadreza Amirian</td>
<td>Doctoral Student</td>
<td><a href="mailto:amir@zhaw.ch">amir@zhaw.ch</a></td>
</tr>
<tr>
<td>Regula Chase</td>
<td>Administrative Assistant</td>
<td><a href="mailto:chas@zhaw.ch">chas@zhaw.ch</a></td>
</tr>
<tr>
<td>Dr. Ricardo Chavarriaga</td>
<td>Senior Researcher, MSc programme coordinator</td>
<td><a href="mailto:char@zhaw.ch">char@zhaw.ch</a></td>
</tr>
<tr>
<td>Prof. Dr. Mark Cieliebak</td>
<td>Senior Lecturer, head of NLP Group</td>
<td><a href="mailto:ciel@zhaw.ch">ciel@zhaw.ch</a></td>
</tr>
<tr>
<td>Dr. Jan Milan Deriu</td>
<td>Research Associate</td>
<td><a href="mailto:deri@zhaw.ch">deri@zhaw.ch</a></td>
</tr>
<tr>
<td>Raphael Emberger</td>
<td>Research Assistant</td>
<td><a href="mailto:embe@zhaw.ch">embe@zhaw.ch</a></td>
</tr>
<tr>
<td>Adhiraj Ghosh</td>
<td>Research Intern</td>
<td><a href="mailto:ggos@zhaw.ch">ggos@zhaw.ch</a></td>
</tr>
<tr>
<td>Nicola Good</td>
<td>Research Assistant, Master student</td>
<td><a href="mailto:goon@zhaw.ch">goon@zhaw.ch</a></td>
</tr>
<tr>
<td>Manuela Hürlimann</td>
<td>Research Associate</td>
<td><a href="mailto:huer@zhaw.ch">huer@zhaw.ch</a></td>
</tr>
<tr>
<td>Nikolaos Kapotis</td>
<td>Research Associate</td>
<td><a href="mailto:kapt@zhaw.ch">kapt@zhaw.ch</a></td>
</tr>
<tr>
<td>Katsiaryna Mlynchyk</td>
<td>Research Assistant</td>
<td><a href="mailto:mlyn@zhaw.ch">mlyn@zhaw.ch</a></td>
</tr>
<tr>
<td>Dr. Javier Montoya</td>
<td>Research Associate</td>
<td><a href="mailto:mony@zhaw.ch">mony@zhaw.ch</a></td>
</tr>
<tr>
<td>Daniel Sebastian Neururer</td>
<td>Research Assistant</td>
<td><a href="mailto:neud@zhaw.ch">neud@zhaw.ch</a></td>
</tr>
<tr>
<td>Pascal Sager</td>
<td>Research Assistant, Master student</td>
<td><a href="mailto:sage@zhaw.ch">sage@zhaw.ch</a></td>
</tr>
<tr>
<td>Dr. Frank-Peter Schilling</td>
<td>Senior Researcher</td>
<td><a href="mailto:scin@zhaw.ch">scin@zhaw.ch</a></td>
</tr>
<tr>
<td>Zhivar Sourati Hassanzadeh</td>
<td>Research Intern</td>
<td><a href="mailto:soua@zhaw.ch">soua@zhaw.ch</a></td>
</tr>
<tr>
<td>Prof. Dr. Thilo Stadelmann</td>
<td>Senior Lecturer, head of CVPC Group, director of Centre</td>
<td><a href="mailto:stdm@zhaw.ch">stdm@zhaw.ch</a></td>
</tr>
<tr>
<td>Dr. Don Tuggener</td>
<td>Senior Researcher</td>
<td><a href="mailto:tuge@zhaw.ch">tuge@zhaw.ch</a></td>
</tr>
<tr>
<td>Lukas Tuggener</td>
<td>Doctoral Student</td>
<td><a href="mailto:tugg@zhaw.ch">tugg@zhaw.ch</a></td>
</tr>
<tr>
<td>Pius von Däniken</td>
<td>Research Assistant, Master student</td>
<td><a href="mailto:vode@zhaw.ch">vode@zhaw.ch</a></td>
</tr>
<tr>
<td>Christoph von der Malsburg</td>
<td>Visiting Professor</td>
<td><a href="mailto:xvne@zhaw.ch">xvne@zhaw.ch</a></td>
</tr>
</tbody>
</table>
A.3 Location

CAI Centre for Artificial Intelligence
Technikumstrasse 71
PO Box
CH-8401 Winterthur
www.zhaw.ch/en/engineering/institutes-centres/cai/

Contact
Prof. Thilo Stadelmann
Phone +41 58 934 72 08
thilo.stadelmann@zhaw.ch

Administration
Regula Chase
Phone +41 58 934 76 48
office.cai@zhaw.ch
School of Engineering

CAI Centre for Artificial Intelligence

Technikumstrasse 71
PO Box
CH-8401 Winterthur
www.zhaw.ch/cai/en

Contact
Prof. Thilo Stadelmann
+41 58 934 72 08
thilo.stadelmann@zhaw.ch

Administration
Regula Chase
+41 58 934 76 48
office.cai@zhaw.ch