Subject- and task-independent neural correlates and prediction of decision confidence in perceptual decision making

Journal of Neural Engineering

PAPER

Subject- and task-independent neural correlates and prediction of decision confidence in perceptual decision making

Jacobo Fernandez-Vargas[1], Christoph Tremmel[1], Davide Valeriani[2,3], Saugat Bhattacharyya[1,4], Caterina Cinel[1], Luca Citi[1] and Riccardo Poli[1,*]

[1] Brain-Computer Interfaces and Neural Engineering laboratory, School of Computer Science and Electronic Engineering, University of Essex, Essex, United Kingdom
[2] Department of Otolaryngology | Head and Neck Surgery, Massachusetts Eye and Ear, Boston, MA, United States of America
[3] Department of Otolaryngology | Head and Neck Surgery, Harvard Medical School, Boston, MA, United States of America
[4] School of Computing, Engineering & Intelligent Systems, Ulster University, Londonderry, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: rpoli@essex.ac.uk

Keywords: BCI, confidence, EEG, decision-making, transfer-learning, neural correlate

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
21 October 2020

REVISED
27 January 2021

ACCEPTED FOR PUBLICATION
29 March 2021

PUBLISHED
13 May 2021

Abstract

Objective. In many real-world decision tasks, the information available to the decision maker is incomplete. To account for this uncertainty, we associate a degree of confidence to every decision, representing the likelihood of that decision being correct. In this study, we analyse electroencephalography (EEG) data from 68 participants undertaking eight different perceptual decision-making experiments. Our goals are to investigate (1) whether subject- and task-independent neural correlates of decision confidence exist, and (2) to what degree it is possible to build brain computer interfaces that can estimate confidence on a trial-by-trial basis. The experiments cover a wide range of perceptual tasks, which allowed to separate the task-related, decision-making features from the task-independent ones. *Approach.* Our systems train artificial neural networks to predict the confidence in each decision from EEG data and response times. We compare the decoding performance with three training approaches: (1) single subject, where both training and testing data were acquired from the same person; (2) multi-subject, where all the data pertained to the same task, but the training and testing data came from different users; and (3) multi-task, where the training and testing data came from different tasks and subjects. Finally, we validated our multi-task approach using data from two additional experiments, in which confidence was not reported. *Main results.* We found significant differences in the EEG data for different confidence levels in both stimulus-locked and response-locked epochs. All our approaches were able to predict the confidence between 15% and 35% better than the corresponding reference baselines. *Significance.* Our results suggest that confidence in perceptual decision making tasks could be reconstructed from neural signals even when using transfer learning approaches. These confidence estimates are based on the decision-making process rather than just the confidence-reporting process.

1. Introduction

1.1. Decision-making

A decision is the result of a process that integrates contextual cues and pre-existing knowledge to commit to a categorical choice to achieve a particular goal. It has been shown that during the decision-making process the human brain weighs and integrates multiple noisy sources of information over time [1–5]. As a result, a meta-cognitive evaluation of the decision is generated: the confidence [6–8], which reflects the perceived probability of being correct and is generally correlated with the accuracy, similarly to other behavioural and physiological measures, such as the response time (RT) [9–15].

Different neural correlates of decision making have been identified using neuroimaging techniques such as electroencephalography (EEG) [16, 17], including neural correlates of confidence [8, 11, 18–20]. In particular, the activity in the pre-frontal [21, 22] and parietal [23, 24] cortices correlates with the confidence reported by human participants.

EEG has been used to characterise numerous brain states, such as mental workload, valence, and arousal, which directly or indirectly affect decision making [25–27]. Also, several studies have found differences in the event-related potentials (ERP) for different confidence levels in decision making [5, 18, 20, 23, 28, 29]. These differences in the brain activity make it possible to predict and classify confidence on a decision-by-decision basis using machine learning algorithms [7, 20, 30, 31].

1.2. Brain–computer interfaces

Significant experience in the detection, prediction, and classification of trial-by-trial brain responses has been acquired in the field of brain–computer interfaces (BCIs). Normally, these devices are used to create a communication channel between a human with significant motor disabilities and a machine [32]. The most common BCIs record brain activity via EEG, thanks to its low cost, high temporal resolution, non-invasiveness, and practicality. The downside of EEG is the low signal-to-noise ratio, non-stationarity, and low spatial resolution. Due to these limitations, standard BCIs can only issue a small set of commands and do so rather slowly and with occasional errors, although the trade-off between speed and accuracy can often be adjusted [33–35].

BCIs can also be used for other forms of cognitive human augmentation [36]. For instance, if EEG signals are complemented with behavioural and other physiological recordings (hybrid BCIs), one can obtain systems to support group decision-making [37–41] that are capable of delivering significant practical benefits in real-world situations.

1.3. Transfer learning

Due to high inter-subject variability of EEG signals, BCIs are usually required to be trained with the brain signals of each user. The length of the training process limits BCI applicability in many domains. Transfer learning could significantly reduce the training needs of a BCI, by training the machine-learning model of the BCI with data from one participant, and use the trained model with a different participant [42].

To date, transfer learning has only been used in established BCI applications, such as ERP detection [43, 44], motor imagery classification [45, 46], and steady state visual evoked potentials [43, 47]. This is, mainly, because of the non-stationarity of EEG, that limits the efficiency of such approach. To the best of our knowledge, transfer learning has not yet been applied to confidence decoding in decision making.

1.4. Contributions

This paper makes the following contributions.

1.4.1. Neural correlates of confidence

Most studies reporting analysis techniques for confidence do it only for one task [4, 6, 7, 12, 20, 21, 23, 31, 48]. In this study, we instead investigate the neural correlates of the confidence across *eight* different experiments with 68 participants in total. This is a significative increase compared to classical BCI experiments where 5–20 participants are used. In most of the cited papers, the confidence levels are divided in two groups (Confident and non-Confident), either by actively asking the participant only those two options, or calculating the median. We grouped the confidence into four levels (low, mid, high, and sure) to find whether there was a gradient in the neural correlates associated to the confidence level. All experiments included decision tasks where information was presented visually, but had different stimuli and feedback. Analysing multiple experiments allowed to increase the generality of results and interpretations. With this approach we expected to find both tasks-related differences in the processes associated with confidence evaluation, as well as a task-independent common biomarker of confidence. To the best of our knowledge this is the first time an approach like this is been attempted.

1.4.2. Confidence prediction

The second goal of this study was to predict the reported confidence. Of course, the best way to obtain the confidence is not to predict it but ask directly the participant after each decision. However, being able to accurately predict decision confidence is important in many time-critical tasks (e.g. in the military, in trading, etc) where waiting for people to evaluate and express their confidence is not viable.

In previous studies, we focused on predicting the probability of each trial being correct, as this is particularly useful for aiding group decisions [38–40, 49–51]. In this study however, we focus our attention on *predicting the confidence reported by participants* after each decision. The rationale behind this choice is that, after task familiarisation, reported confidence tends to be a good estimator of performance [29]. Another advantage of estimating confidence rather than probability of correctness is that the latter may not be readily available in every task to be able to train the BCI, while the former can be asked at any time to the user during the training period.

Because confidence is essentially an analogue quantity, *we treat confidence prediction as a regression problem with analogue outputs*, and do not require any prior knowledge of the participant's confidence distribution compared to more traditional approaches where the confidence prediction is treated as a binary problem [7, 20, 31]. BCI systems are rarely used to solve regression problems. However, there do exist

some prior examples. For instance, the estimation of drowsiness [52], reaction time [53] or hand position [54]. Nevertheless, to the best of our knowledge, a regression approach has never been attempted for the prediction of the reported confidence.

Finally, for the first time, we investigated a zero-training approach to confidence prediction. Zero-training is a form of transfer learning where the predictor is not tailored for each participant. We used different approaches to predict the confidence to investigate their impact on the quality of the prediction.

1.4.3. Validation

We validated our BCI confidence decoders using general zero training in experiments where the participants did not report their confidence in the decisions. This allowed us to investigate whether the neural correlates of confidence identified in this study were related to the whole decision-making process, rather than being neural correlates of confidence reporting.

2. Materials and methods

2.1. Experiments

For this study, eight different experiments conducted over the past 6 years were used. As it can be seen in figure 1, all experiments had similar trial structures: visual information was provided, as either a static image or a video sequence. Following which, a decision had to be made and reported by the participant together with their confidence (except for the two validation experiments).

A brief description of each experiment is given below, while full details of each experimental protocol are included in the corresponding publications.

2.1.1. PATROL1 [50], PATROL2, and PATROL3

In these three experiments, participants were presented with a video of a corridor with doors at both sides. At a random time, a soldier figure appeared for 250 ms, and participants had to decide (within a 2 s timeout) whether the figure was wearing a helmet or a cap, by pressing the left or right mouse button, respectively. Participants were then asked to report the confidence in their decision using the mouse wheel. A blue bar representing the confidence varied accordingly as participants manipulated the wheel. If the participant did not respond, the experiment would continue and their decision would be considered incorrect (miss). In PATROL2 and PATROL3, after confidence was provided, feedback on the correctness of the decision and confidence assessment was given to participants. In PATROL2 feedback was in the form of a slider, which was most negative for incorrect fully-confident decisions, weakly negative for incorrect low-confidence decisions, weakly positive for correct low-confidence decisions, and most positive for correct fully-confident decisions. In PATROL3, participants were shown their decision (represented by the labels 'Cap' and 'Helmet') and confidence (represented by a confidence slider) side by side with the decision and confidence of an expert. Twelve participants took part in PATROL1 experiment, while ten participants undertook PATROL2 and PATROL3. Each participant performed 336 trials.

2.1.2. ASSISTED-PATROL

This experiment, while being similar to the PATROL experiments, presented some differences. Firstly, instead of video feeds, two static images were presented (see figure 1): (1) the empty corridor and (2) the corridor with a character, both displayed for 250 ms. Secondly, after the character image, participants were forced to respond (i.e. there was no timeout for their responses). Thirdly, some of the trials contained a cue before the stimulus indicating if the target would appear to the right- or to the left-hand-side of the corridor. Finally, there was no trial-by-trial feedback. There were four types of trials: without any cue, with a static head that did not provide any information, a voice saying 'right' or 'left', or a talking head saying either 'right' or 'left'. In this experiment, 12 participants were tested, each performing 640 trials, 160 per condition[5].

2.1.3. SUSPECT-DETECTION [40]

In each trial of this experiment, participants were presented with a black-and-white photo of a crowded corridor (for 300 ms), and then with a display asking if a specific person was present. They had to answer yes or no by pressing the left or right mouse buttons, respectively, and then had to report their confidence using the mouse wheel as in the PATROL experiments. In this experiment, 10 participants were tested, each performing 288 trials.

2.1.4. OUTPOST [51]

In this experiment, participants were shown a video sequence simulating the viewpoint of soldier at an outpost surveying a clearing. In each trial, a character appeared in the distance and walked towards the outpost. Participants were asked to decide as quickly as possible whether the character was wearing a helmet or a cap, reporting their decision using the mouse buttons. They then had to report the confidence in this decision using the mouse wheel. The character

[5] The experiment was run (with the same protocol and amplifiers) jointly at the University of Essex and at the University of Southern California. EEG was recorded from six participants (tested at Essex) using a 64-electrode cap, while for the other six a 256-electrode EEG cap was used. The 256-electrode EEG dataset was down-sampled with bilinear interpolation to reconstruct EEG signals from similar locations as for the 64-electrode cap. For four electrode locations (AF7, P9, AF8 and P10) this was not possible, and the signals from these electrodes were discarded for all 12 participants. Hence, preprocessing and data analysis for ASSISTED-PATROL were done using 60 electrodes.

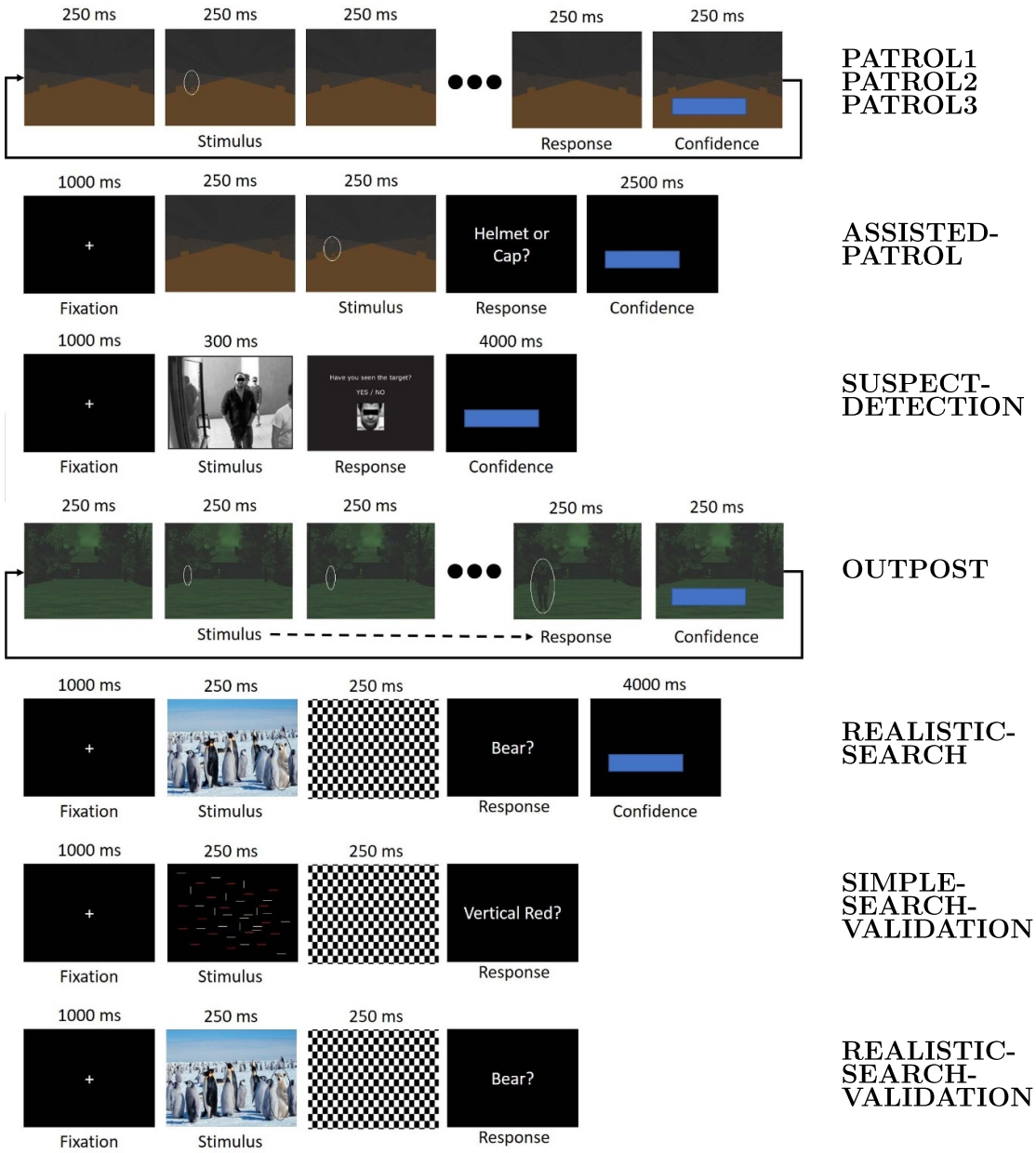

Figure 1. Trial structure for the experiments used in this article. In PATROL and OUTPOST, the arrow from the last to the first frame indicates that trials form a continuous sequence with no pauses, while the three black dots indicate that the system shows one frame every 250 ms (4 Hz). The dashed arrow in OUTPOST indicates that the stimulus changes over time. Finally, ellipses were superimposed to help the reader identify the character.

stayed on screen until the participant responded, so stimulus presentation time was not fixed. However, there was a variable RT to make this decision. In this experiment 10 participants were tested, each doing 360 trials.

2.1.5. REALISTIC-SEARCH1 and REALISTIC-SEARCH2 [39]

In these two experiments, participants had to perform a visual search task. An image of an arctic environment with a variable number of penguins and possibly a polar bear, photorealistically imposed on the image, was presented for 250 ms. Participants had to decide whether there was a polar bear or not in the image. Response and confidence were expressed in the same way as in previous experiments. The difference between REALISTIC-SEARCH1 and REALISTIC-SEARCH2 was that, in the latter, after reporting their confidence, participants were informed about the confidence and decisions reported by another participant, and had the possibility of changing their response. In the REALISTIC-SEARCH1 experiment 10 participants were tested, while 16 took part in the REALISTIC-SEARCH2 experiment. Each participant performed 320 trials.

2.1.6. SIMPLE-SEARCH-VALIDATION [38]

This experiment was composed of 320 trials, where participants were presented with a display composed

of 40 green and red lines, either vertical or horizontal, on a black background for 250 ms. Their task was to decide whether or not there was a vertical red bar in the image by pressing the left or right mouse buttons. Ten people took part in the experiment. This experiment was only used to validate the models because the confidence was not reported.

2.1.7. REALISTIC-SEARCH-VALIDATION [49]

The task was the same as for REALISTIC-SEARCH1, except that participants were not asked to report their confidence. Ten subjects took part in the experiment, each performing 320 trials. This experiment was not used for training, but only to validate the models in situations where confidence was not reported.

In all experiments, confidence was reported using a scale from 0 to 1 in steps of 0.1, creating 11 possible reported confidence values.

2.2. Exclusion criteria

Both the raw signals and the descriptive statistics of the EEG recordings were inspected to determine the quality of the recordings. This resulted in the removal of three participants from PATROL1 and one from PATROL2. Furthermore, we excluded the participants with an accuracy more than two standard deviations lower than the average accuracy across all participants. These participants were likely not paying attention to the experiment or did not understand the task. As a result, six participants with accuracy lower than 56% were excluded: two from SUSPECT-DETECTION, one from REALISTIC-SEARCH1, and three from REALISTIC-SEARCH2. Moreover, because we were focusing our analysis on the confidence, we removed those participants for which the distributions of confidence in the correct trials and in the incorrect trials were not significantly different, as assessed by the Wilcoxon rank-sum test resulting in $p > 0.05$. After this process another 12 participants were excluded: one from PATROL2, one from SUSPECT-DETECTION, three from OUTPOST, one from REALISTIC-SEARCH1, and six from REALISTIC-SEARCH2. Therefore, data from a total of 68 participants were included in the analysis.

We also removed the trials in which the participants reported a 0 confidence, as this measure was used by participants in different ways. Some would use it to indicate that they had responded randomly, while others used it to indicate a wrong decision. In total, these trials represented only 1.7% of the total. Full details about the distribution of the accuracy, RT, and proportion of each confidence level can be found in tables S1–S3 respectively in the additional materials (available online at stacks.iop.org/JNE/18/046055/mmedia).

2.3. Setup and preprocessing

In all experiments, participants sat comfortably at about 80 cm from an LCD screen while wearing an EEG cap connected to a Biosemi ActiveTwo system. All the experiments were performed using wet electrodes and, except for six participants in ASSISTED-PATROL, all the recordings were performed with 64 electrodes in the standard 10–20 system.

The EEG data were preprocessed as described in [39]. In brief, the original data was sampled at 2048 Hz, then band-pass filtered between 0.15 and 40 Hz using an FIR filter. Then, the signal was down sampled by a factor of 16, resulting in a 128 Hz signal. In addition to this, a correction for eye-blink and other ocular movements was performed using a subtraction algorithm based on correlations to the average differences between FP1 and F1 and between Fp2 and F2 [55].

After the preprocessing, two types of epochs were extracted from EEG in each trial: stimulus-locked and response-locked. The former started at stimulus onset and lasted for 2.5 s, to ensure inclusion of the response and its neural correlates in every experiment. Response-locked epochs started 1.25 s before the response and lasted for 1.5 s, as we were mostly interested into the neural processes leading to a decision. A baseline was calculated and removed for each epoch and channel. In each epoch, the mean of the signal from 25 ms before and 25 ms after the stimulus or response onset was used as baseline correction. Finally, when performing ERP analyses, an epoch rejection process was applied. For each participant, the difference between the maximum and minimum voltage of each epoch was computed. Then, a threshold was set as the third quartile plus 1.5 times the difference between the first and the third quartiles. The epochs that had a difference between the maximum and the minimum voltage higher than the threshold were removed from the analysis, on a channel-by-channel basis.

2.4. Neural correlates of decision confidence

We grouped trials into four categories: *Low confidence*, for trials with a reported confidence between 0.1 and 0.3; *Mid confidence*, for trials with a reported confidence between 0.4 and 0.6; *High confidence*, for trials with a reported confidence between 0.7 and 0.9; and *Sure*, for trials with a reported confidence of 1. The reason to have an independent category for *Sure* is that participants report full confidence much more frequently than any other value of confidence (32% of the trials overall). We chose fixed boundaries for the confidence compared to other studies because this allowed use of the same methodology across participants and tasks. Having boundaries based on the percentiles of the reported confidence allows to have a somewhat calibrated confidence, as well to have

a more balanced set of labels for the classification. However, to do this its necessary to have the distribution of the confidence of the user a-priori.

Two methods were used to analyse the EEG activity for different confidence levels. First, we performed a *single-experiment analysis*, where we calculated the average epoch voltage across trials for each confidence level (without grouping by subject), as well as the 95% confidence interval.

Next we performed a *multi-task analysis*, where we first calculated the epoch average of each participant and confidence level. Then the grand average from those averages was derived. Participants that did not have at least 10% of trials at each confidence level were removed from this analysis. This resulted in three participants being removed: one from SUSPECT-DETECTION and two from REALISTIC-SEARCH2.

Statistical analyses were conducted using the non-parametric Kruskal–Wallis test comparing the average voltages between confidence levels for each time point and channel.

2.5. Confidence regression and transfer learning

The system was designed with the following structure:

(a) Feature extraction: uses the raw EEG data and RT as input for each trial (32 769 values) and returns the calculated features (129).
(b) Feature selection: using a greedy algorithm based on linear regression, the features that do not contribute positively to the prediction are removed.
(c) Predictor: using the selected features, an artificial neural network (ANN) with the same number of inputs as the number of features selected, and 10 neurons in the output layer is used. The activations of the output layer are weighted to obtain the final prediction.

In this section we will describe each of these items one by one. The three approaches used for different levels of transfer learning were implemented following the upcoming structure.

2.5.1. Features

Two features for each EEG channel were selected from each trial, one from the stimulus-locked epoch and another from the response-locked epoch. For the stimulus-locked epochs we used the ERP amplitude, calculated as the mean voltage between 500 and 750 ms after stimulus onset minus the mean voltage in the preceding 500 ms. For the response-locked epochs, we used the mean amplitude of the EEG between 1250 and 500 ms before the response. Combining these features for all channels we had 128 neural features (120 for the ASSISTED-PATROL experiment) for each trial.

Additionally, we used the RT as a feature, since it has been demonstrated that both confidence and correctness are correlated with RT in a variety of situations [13, 14, 24, 37, 56, 57].

To investigate the contribution that the RT may have over the prediction, we performed the training and testing using only the RT as an input for the classifier.

2.5.2. Feature selection

The method used to select the features was an iterative greedy algorithm that first measured the prediction accuracy with a set of features. Then it removed one feature at a time, re-calculating the prediction accuracy to determine how much this changed when the feature was removed. Then the algorithm permanently discarded the two features that, when removed, changed prediction accuracy the least. At that point the algorithm started again, repeating the process with the remaining features until the desired number of features remained. This algorithm is an adaptation of an algorithm described in previous work [58]. Here, when we calculated prediction accuracy, we trained and tested (with five-fold cross-validation) a linear regression model using the available features. We used linear regression instead of an ANN for reducing the computational burden of this optimisation procedure.

2.5.3. Predictor

The first step of the prediction was an ANN that followed a shallow network approach with 10 neurons in the output layer. The ANN was designed to predict the confidence from the selected features with the aforementioned algorithm. This meant that different models had different input sizes. Each feature was z-scored corrected, as this is standard for most machine learning systems. This method prevents features with a higher mean and variation from becoming more relevant to the classification than they should be. The network had 10 output neurons, one for each of the ten possible confidence levels (from 0.1 to 1 in steps of 0.1). A dropout layer (with a probability of 0.5) was used to reduce overfitting [59]. The network had one hidden layer, the neurons of which used a hyperbolic-tangent activation function. Finally, a softmax layer was used to sharpen the network outputs.

Cross validation was used to decide the number of neurons in the hidden layer. The values tested were 5, 10 and 20. The maximum number of training epochs was set to 500, with a mini batch size of 10% of the data. In addition to this, an early-stop criterion based on the error of a validation set was used. A fraction (20%) of the trials were extracted, without replacement, from the training set to form the validation set. Training was stopped after six epochs in which the error on the validation set did not decrease.

The ith output of the softmax layer for a particular input pattern was taken to represent the probability, w_i, of confidence level $cl_i \in \{0.1, 0.2, \ldots, 1\}$ being reported in a trial, for all i's. However, instead of using the most probable confidence level as the predicted confidence, we used a weighted sum of the most probable confidence levels. The goal of this was to obtain an analogue output, not limited to just 10 different confidence values. To predict the confidence in a trial, the N most probable confidence levels cl_i's were selected. Assuming the reorder the softmax outputs by w_i, w_1 being the largest, w_2 the second largest and so on, N was set trial by trial to be the smallest number for which $\sum_{i=1}^{N} w_i \geqslant 0.5$[6].

Then, the predicted confidence for that trial was calculated using the following formula:

$$\frac{\sum_{i=1}^{N} w_i * cl_i}{\sum_{i=1}^{N} w_i}. \quad (1)$$

2.5.4. Prediction approaches

To assess the efficacy of a zero-training method, we tested three different approaches. These approaches were validated using cross validation, while the cross validation used for the feature selection was nested inside.

2.5.4.1. Single subject (SS)

A model was trained and tested for each participant individually, using standard cross validation. This non-zero-training approach provided a baseline reference level of performance to compare the other two approaches.

2.5.4.2. Multi subject (MS)

A transfer-learning model across subjects was trained and tested separately for each experiment, using a *leave-one-subject-out* cross-validation approach. The results from this model represented the performance of a system that does not know anything about the future user, but is likely specialised to the task performed in each experiment. This approach was a form of transfer learning across participants.

2.5.4.3. Multi task (MT)

In this generalised model with transfer learning across tasks, the training and test sets followed a *leave-one-experiment-out* cross-validation approach. This model represented a fully-generalised approach, where the system is independent from both the subject and the perceptual decision-making task.

[6] Setting this threshold to 0.5 was a compromise between the need to provide a more graded confidence prediction than using cl_1 (i.e. the most probable output of the ANN) and the need to be able to predict values at the extreme of the range of the confidence scale (e.g. 1, which is the most common response).

2.5.5. Evaluation methods

For each of these approaches, we used an ANN as predictor (details are in a later section). As a baseline for the ANN we selected the prediction error made by a classifier that always predicted the mean of the reported confidence (this baseline is more conservative than using just random). To evaluate the different approaches we used the following metrics: (1) the median absolute error (MAE) between the predicted confidence and the reported confidence (MAE), (2) the prediction's median meta-cognitive accuracy (MCA), (3) the prediction's mean confidence delta ($c\Delta$), and (4) the prediction's mean confidence calibration. We defined (2)–(4) below.

The *meta-cognitive accuracy* (MCA) is a quantity that indicates how good a person is at evaluating their own decisions [60]. There are different methods of measuring it [29, 61] with their own advantages and disadvantages. Here, we computed the MCA as follows:

$$\begin{aligned} \text{MCA} &= 1 - |\text{confidence} - \text{correctness}| \\ &= \begin{cases} \text{confidence} & \text{if correctness} = 1, \\ 1 - \text{confidence} & \text{otherwise}, \end{cases} \end{aligned} \quad (2)$$

where the correctness is 0 (incorrect) or 1 (correct), and the confidence is a value between 0 and 1. Therefore, the MCA takes values between 0 and 1, with 1 representing a correct decision made with the highest confidence or an incorrect decision made with 0 confidence, and 0 representing an incorrect decision made with the highest confidence or an correct decision made with 0 confidence.

The formula above applies to each individual decision. We were however, interested in aggregate statistics, such as $E[MCA]$ for a participant across all trials in an experiment.

The limitations with this measure are that the MCA is heavily influenced by the difficulty of the task and may be influenced in a counter intuitive manner by confidence biases (such as those seen in over confident or under confident individuals). For instance, in a relatively easy task (where participants make infrequent mistakes) a participant could obtain a high average MCA by just responding always with the highest value of confidence, irrespective of whether the decision was correct or incorrect.

To complement the MCA, we also defined a new measure, the *confidence delta* or $c\Delta$, which represents to what extent the confidence recorded in correct decisions is higher than that recorded in incorrect decisions:

$$\begin{aligned} c\Delta = E[&\text{confidence}|\text{correctness} = 1] \\ &- E[\text{confidence}|\text{correctness} = 0]. \end{aligned} \quad (3)$$

A value of $c\Delta = 1$ would indicate that the participant reported full confidence to all the correct trials and

a 0 confidence to all the incorrect ones. A $c\Delta = 0$ would indicate that reported confidence is random or that the same confidence is reported for every trial. We calculated this value on a subject-by-subject basis.

Finally, another desirable property of the confidence (either reported or predicted) is *calibration*. Confidence is calibrated if, on average, it matches the probability of decisions being correct. To quantify the degree of calibration we defined the following *calibration offset*:

$$\text{calibration offset} = \big|E[\text{confidence}] - P(\text{correctness} = 1)\big|. \quad (4)$$

This value reflects how close the predicted confidence is to the accuracy of the participant, a positive *calibration offset* indicating for instance that a participant is either overconfident or under-confident.

When taken together, these metrics provide a comprehensive evaluation on the quality of the our confidence predictors.

2.5.6. Statistical analysis

For each one of the four evaluation methods, first we performed a Wilcoxon test to compare the ANN and the Baseline methods. Then, we performed a Kruskal–Wallis analysis using only the ANN methods. If this resulted in a p-value < 0.05, we then performed three Wilcoxon pair test.

2.6. Model validation

As mentioned before, four measures were used to assess the quality of the prediction: MAE, MCA, $c\Delta$ and calibration offset.

To further validate the MT models, we used two additional experiments (SIMPLE-SEARCH-VALIDATION and REALISTIC-SEARCH-VALIDATION; see 'Experiments' section and figure 1) as validation data. The confidence for those experiments was predicted using the trained models. The difference with the training experiments was that participants were not asked to report their confidence in the validation experiments.

To predict the confidence in these two experiments, a new ANN was trained with the MT approach, but cross validation was not used. Instead all eight experiments were used for training, using the number of neurons in the hidden layer that resulted in the lowest error during the training of the MT models.

To evaluate performance in the validation experiments, the MCA, $c\Delta$, and calibration offset were calculated using the outputs of the ANN on the two validation experiments. It was not possible to calculate the MAE given that there was no ground truth to compare the predictions with. This validation method allowed us to test whether the neural correlates observed and, thus the prediction made, arose from the decision-making process or were linked to the confidence-reporting process.

3. Results

In this section, the results of the study are reported. These are divided into three areas:

Neural correlates of confidence decision, in which we analysed the neural correlates of confidence in a large dataset of eight different experiments and 68 participants. Furthermore, we divided the confidence level into four groups to get a finer grain analysis compared to the typical confident vs non-confident categorisation.

Confidence prediction and transfer learning, where we used a BCI system to predict the confidence in single trials. Using BCIs to predict continuous values is not common; in particular, we could not find any study predicting confidence in decision-making. In addition, we investigated the feasibility of using transfer learning to build a system that can be used in a plug-and-play manner.

Model validation, where, exploiting the transfer leaning capabilities of our system, we went a step further and predicted the confidence of decision-making tasks where participants were not asked to report their confidence. This validation provides further support for the results and shows how generalizable the proposed method is.

3.1. Neural correlates of decision confidence

3.1.1. ERPs

Figure 2 shows the results for stimulus-locked epochs and channel Pz as the parietal area has previously shown to correlate with the confidence during perceptual decision making [23]. In most experiments we see differences in the averages recorded for different confidence levels at 350–600 ms after the stimulus onset. Also, ERPs recorded in the 'Sure' condition were, in most tasks, higher than in the other levels of confidence. The PATROL1 and OUTPOST experiments were exceptions to this behaviour. In the case of the OUTPOST experiment, we could not see any real ERPs. This is because, due to the nature of the experiment design, the appearance of the character on the display can be quite difficult to detect at first, and participants might take a relatively long time from the appearance of the stimulus before they are ready to respond. Hence, there is a large variability in ERPs latencies across trials and participants, resulting in flat averages. In the case of the PATROL1 experiment, ERPs were present but the stimulus-locked epochs have no difference between the confidence bins.

Figure 3 shows the results for response-locked epochs. Here, the differences between the four confidence levels are much more visible, and the ERPs' morphology seems to vary proportionally with the confidence level. Using response-locked

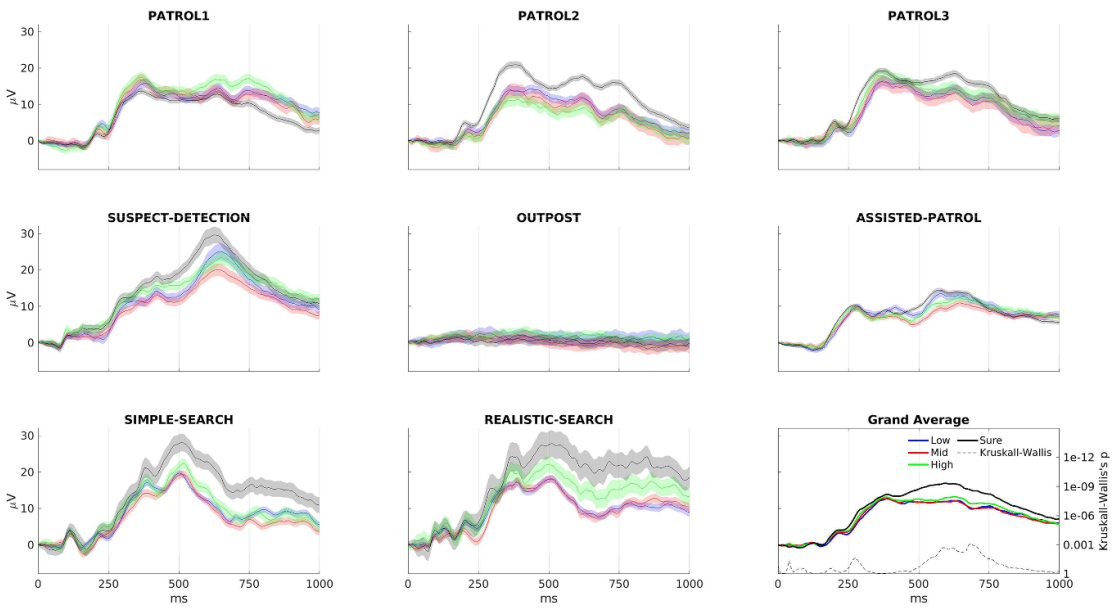

Figure 2. Mean and 95% confidence interval of EEG activity at electrode Pz in *stimulus-locked* epochs for the four confidence levels grouped by experiment. The grand average plot (bottom right) shows the average ERPs across experiments and the *p*-value of the Friedman test (logarithmic ordinate scale on the right-hand side of the plot).

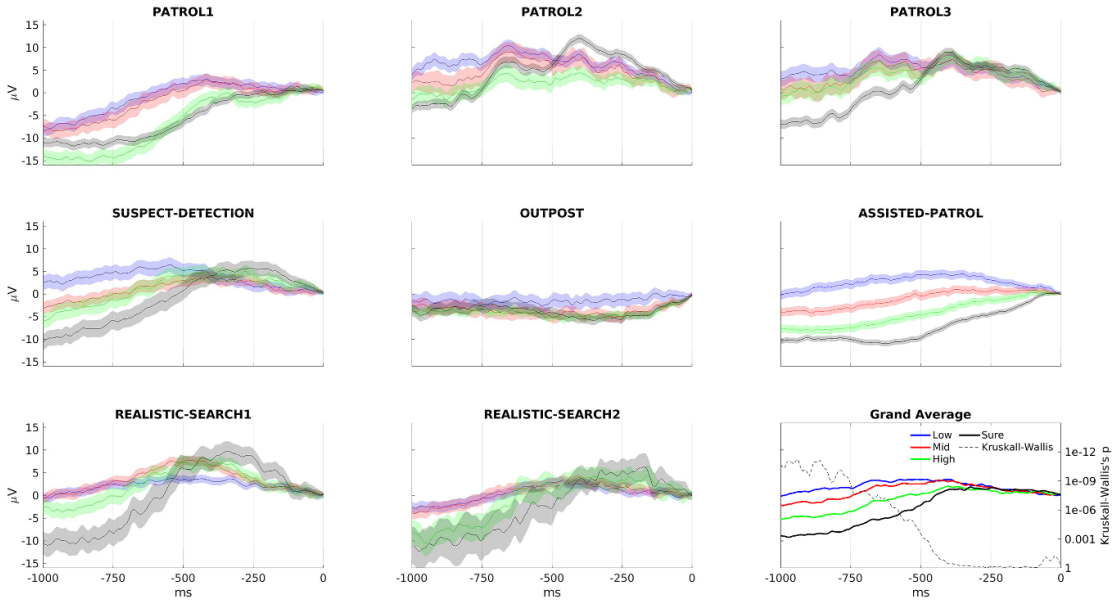

Figure 3. Mean and 95% confidence interval of EEG activity corresponds to the electrode Pz in *response-locked* epochs for the four confidence levels grouped by experiment. The grand average plot (bottom right) shows the average ERPs across experiments and the *p*-value of the Friedman test (logarithmic ordinate scale on the right-hand side of the plot).

epochs, it is possible to observe differences between classes in PATROL1, which were not visible in the stimulus-locked epochs. However, for the OUTPOST experiment we still did not measure any difference between the confidence classes.

3.1.2. Grand averages across experiments

The bottom right plots in figures 2 and 3 show the grand average across all experiments for channel Pz, and the *p*-value of the Friedman test (logarithmic ordinate scale on the right-hand side of the plot) for stimulus- and response-locked epochs, respectively. The 'Sure' class is significantly different from the grand-averages of the other classes for stimulus-locked epochs only, between 600 and 750 ms after the stimulus onset. On the contrary, the response-locked grand-averages show differences between all four confidence levels, particularly up until around 500 ms before the response.

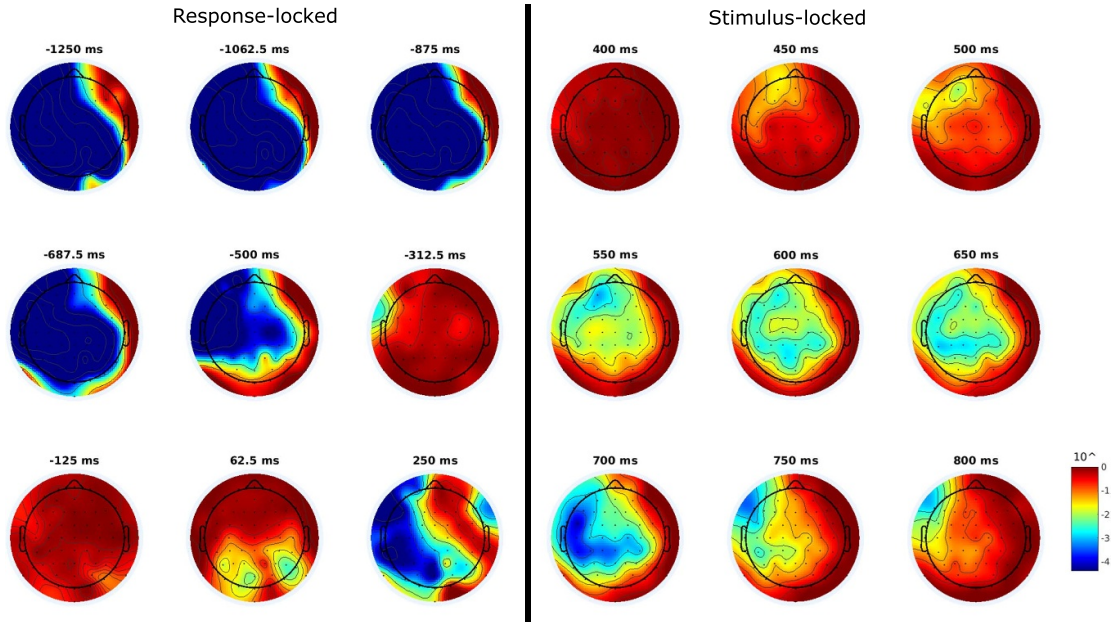

Figure 4. Scalp maps of Kruskal–Wallis's p-values at different times for both stimulus-locked and response-locked epochs. Note that the colour scale is logarithmic. The p-values in the response-locked epochs have been clipped to be comparable to the ones in the stimulus-locked epochs.

Figure 4 shows the Kruskal–Wallis p-values across scalp locations and time for stimulus- and response-locked epochs. The test verifies the hypothesis that ERPs recorded in the four confidence classes are not drawn from the same distribution. Statistical differences among the four classes are lateralised in the left hemisphere, and are present in many channels for both stimulus- and response-locked representations. Differences are stronger between 550 and 750 ms after stimulus presentation in the stimulus-locked epochs. Differences are even stronger in response-locked epochs, as we expected from our earlier observations on the bottom right plots in figures 2 and 3.

To confirm that the differences observed in the grand averages were not a result of differences in correctness rather than confidence levels, we split the data between correct and incorrect trials and performed the same analysis. The result of such analysis can be seen in figure 7 in the additional materials. For correct trials (∼80% of the total) we can observe very similar results to the obtained in the original analysis. This by itself confirms that the differences observed for confidence levels are not due to the correctness of the response. In the case of incorrect trials (∼20% of the total), the differences are only notable for response locked trials, where the effects are magnified by the difference RT between confidence levels. It is important to note that, due to the exclusion criteria, only 32 participants were valid for this analysis.

Finally, to control the effect of the RT, we performed an analysis on the response locked epochs grouping by RT. We created three intervals based on the RT in seconds: [0.5, 0.9), [0.9, 1.2), [1.2, 2). These thresholds were chosen so that, overall, there were the same number of trials for each condition. However, due to the exclusion criteria, this resulted in only 47, 53, and 28 valid subjects being included for each condition, down from the 65 available in the original analysis. The goal was to remove the variance of the RT from the analysis and confirm that the differences in the confidence were not just due to the relative temporal shifts in ERPs induced by response locking. The plots can be seen in figure S8 in the additional materials. The results show that the differences in the response locked epochs are noticeable, in particular for trials falling in the RT intervals [0.5, 0.9) and [0.9, 1.2) which are reasonably narrow. However, we found no differences for trials in the interval [1.2, 2). This confirms that, even if there is a correlation between RT and confidence, not all the confidence variance can be explained by differences in RT.

3.2. Confidence regression and transfer learning

Figure 5 shows the average of the four variables across subjects used to evaluate the quality of the prediction. Furthermore, in figures 9–12 the values for each subject, and task are presented for the MAE, $c\Delta$, and calibration offset respectively. The figure 5(A) shows the MAE between the predicted confidence and the reported confidence across all subjects for the three approaches (SS, MS, and MT) and methods (baseline and ANN) being compared. As expected, the MAE is minimal for the SS approach, slightly worse for MS and even worse for MT. It also appears that baseline is worse than NN.

The MAE of ANN methods was significantly lower than the MAE of the corresponding baseline method (Wilcoxon $p < 0.001$). Moreover, the MAE

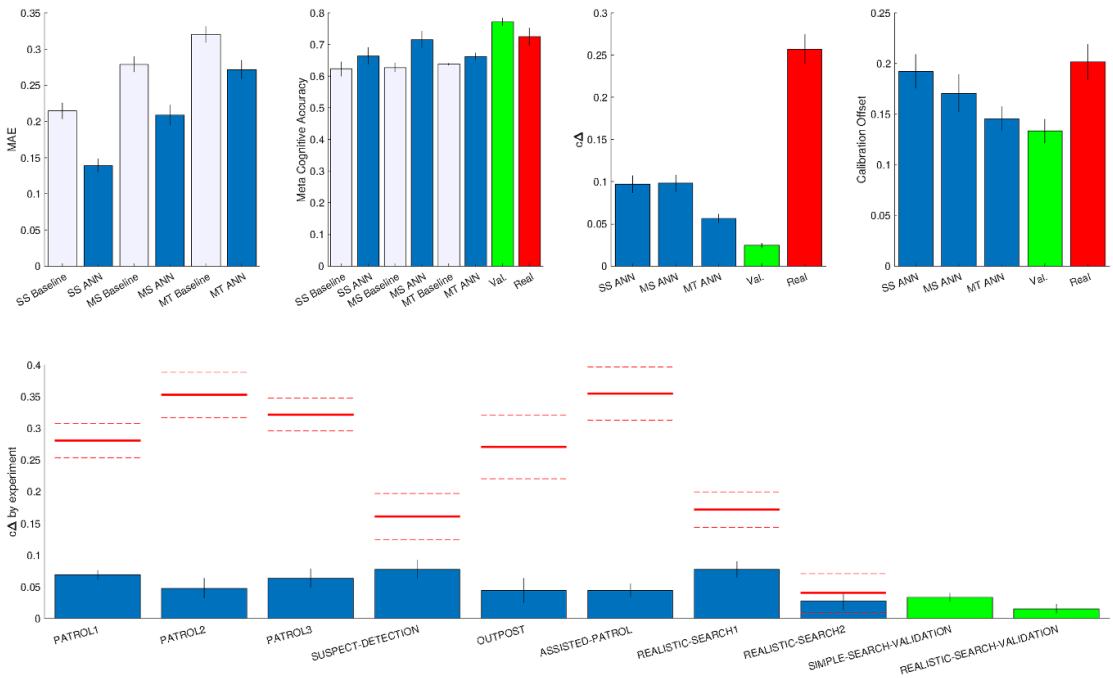

Figure 5. Evaluating the quality of confidence predictions. The four bar charts at the top of the figure show the mean of each measure evaluated across all subjects, while the bar chart at the bottom shows the $c\Delta$ (how different the confidence is between correct and incorrect trials) results for each experiment separately (also averaged across subjects). In the top bar charts, dark blue bars represent the results for the ANN method, light blue bars represent the baseline results, green bars correspond to the validation sets and red bars are the results obtained when the real reported confidence was used. In the bottom plot, red continuous lines indicate the mean of the reported confidence, dotted lines indicate the standard error, and the blue bars show $c\Delta$ for MT predictions. In all bar charts, the vertical black lines indicate standard error of the corresponding data.

of the three ANN-based approaches were significantly different (Kruskal–Wallis $p < 0.001$). Finally, pairwise Wilcoxon rank-sum tests on the ANN-based approaches validated that SS ANN had lower MAE than MS approach, which in turn had lower MAE than the MT approach (Bonferroni-corrected $p < 0.002$).

Additionally, we compared the three approaches when only the RT was used compared to when both the EEG and RT were used. The results showed that using only the RT leads to better results (MAE) for the SS approach (0.109 vs 0.139), but worse for MS (0.213 vs 0.209) and MT (0.282 vs 0.272) approaches. In all cases, the pairwise Wilcoxon rank-sum tests resulted in a Bonferroni-corrected $p < 0.001$. These differences are even more clear if we look at accuracy (weight of the diagonal in the confusion matrix). This value was better for the only RT method in the SS approach (32% vs 36%) but worse for the MS (24% vs 16%) and MT (13% vs 9%) approaches.

Figure 5(B) shows the mean MCA for SS, MS and MT with the Baseline and ANN predictors. The chart also reports the ground truth value ('Real') obtained when using the confidence reported by the participant and the one obtained using the validation data set ('Val.'). We performed the same statistical tests for the MCA as we did for the MAE. All three Wilcoxon rank-sum tests comparing ANN and Baseline for different prediction strategies (SS, MS and MT) resulted in p-values < 0.001, indicating that the ANN achieved a better MCA than the baseline. The Kruskal–Wallis test comparing the three ANN approaches returned a p-value of 0.057, indicating that the three approaches were not statistically different. Because the p-value is (marginally) non significant, we did not perform the corresponding pairwise comparisons. Finally, we further performed three Wilcoxon rank tests to compare the MCA from the ANN (with the SS, MS and MT approaches) to the ground truth value. The Bonferroni-corrected p-values for the three comparisons were 0.095, 1, and 0.029 for SS, MS, and MT, respectively. This indicates that the MCA was significantly worse than the real value only in the MT approach.

Figure 5(C) shows the results for $c\Delta$ for ANN values (as $c\Delta = 0$ for the baseline) as well as ground-truth ('Real') and the validation data set ('Val.'). As it can be seen from the figure, $c\Delta$ of the ANN is much smaller (about 1/4 for MT, and 1/3 for SS and MS) than for the ground-truth (the actual reported confidence). Similarly to the previous performance metrics, we first performed a Kruskal–Wallis test comparing the three ANN approaches. The Kruskal–Wallis test indicated a significant difference between the three ANN approaches ($p = 0.005$). Then, we performed three paired Wilcoxon tests to compare the different approaches. This resulted in the Bonferroni-corrected p-values 1, 0.018, and 0.012 for SS vs MS,

SS vs MT, and MS vs MT, respectively, indicating that the MT approach had a significantly lower $c\Delta$ than the other two methods. Additionally, $c\Delta$ was significantly lower for the reconstructed confidence compared to the ground truth (Wilcoxon $p < 0.001$). Finally, the $c\Delta$ distribution was significantly different from zero ($p < 0.001$ for the three approaches).

Figure 5(E) shows the results of the $c\Delta$ value for individual experiments for the MT approach. This helps us to visualise that, even if the standard error of the $c\Delta$ for the ground-truth is more or less similar across various experiments, the mean varies significantly across them. However, the mean $c\Delta$ of the reconstructed confidence is more stable. It is also interesting to note that there seems to be no correlation between $c\Delta$ values calculated from the ground-truth and the ones calculated from the predicted confidence ($\rho^2 = 0.005$, $p = 0.867$).

Finally, the mean calibration offset is shown in figure 5(D). The confidence reported by participants ('Real') seems less calibrated than the predictions of the ANN, particularly for the MT ANN. As before, the first analysis that we performed was a Kruskal–Wallis analysis to see if there was any difference between the three approaches. This resulted in a p-value of 0.145, suggesting that there was no difference between them. We then compared each of the approaches with the ground truth, which resulted in Bonferroni-corrected p-values of 1, 0.218, and 0.106 for SS, MS, and MT respectively. This indicates that, in terms of calibration offset, the predicted confidence was similar to the real one.

3.3. Model validation

Validation results are represented by the green bars in figure 5. For validation purposes, we reconstructed the (missing) confidence values using an MT ANN approach. As discussed previously, it was not possible to compute MAE between the reported confidence and the reconstructed one because the confidence was not reported in the validation experiments. For each of the three remaining performance metrics, we performed a Wilcoxon rank test to see whether the performance measures obtained with the reconstructed confidence values were statistically different from those obtained with MT ANN approach on the eight training experiments. Both the comparisons of the MCA and $c\Delta$ resulted in $p < 0.001$, while the comparison of the calibration offsets resulted in $p = 0.687$. These results indicate that the validation data performed similarly to the MT ANN approach in terms of calibration offset, better in terms of MCA and worse in terms of $c\Delta$. As with the training data, we performed a Wilcoxon rank-sum test on the $c\Delta$ values to test whether the median was equal to zero. This resulted in a p-value < 0.001. Finally, we tested whether the difference in confidence medians between correct and incorrect responses were statistically significant. This resulted in 48 (70.6%), 48 (70.6%), and 43 (63.2%) participants with significantly different confidence means between correct and incorrect trials for SS, MS, and MT approaches, respectively. Performing the same analysis over the 20 participants of the validation data set, showed that 12 (60%) had a significantly different means.

4. Discussion

In this study, we instead investigate the neural correlates of the confidence across *eight* different experiments with 68 participants in total. Additionally, instead of dividing the confidence into only two groups (confident and non-confident), we grouped the confidence into four levels (low, mid, high, and sure) to find whether there was a gradient in the neural correlates associated to the confidence level.

4.1. Neural correlates of decision confidence

As we have seen in figures 2 and 3 different ERPs are associated with different levels of confidence. This is clearer in response-locked epochs than in stimulus-locked ones. In many practical applications the event that caused a response is not known *a priori*, but the response is, of course, always known. Our results suggest that even if the information about the stimulus onset is not available, it may still possible to obtain strong neural correlates of confidence using response-locked epochs.

Overall, the results indicate that there is a significant degree of similarity in the ERPs associated with different levels of confidence in many experiments. For instance, the higher the confidence the higher the ERP amplitude in several experiments with stimulus-locked epochs and in most experiments with response-locked epochs[7]. In the case of response-locked epochs, this can also be seen in the grand average across experiments on the bottom right of figure 3.

In particular, in several experiments, the condition 'Sure' was associated with significantly bigger ERPs than the other confidence levels in both stimulus- and response-locked epochs. This is visible in both grand averages (in figures 2 and 3). Reporting the confidence requires: deciding the precise confidence to report. This can be easy and quick when the target is comfortably recognised, however, for less confident decisions, deciding the precise confidence to report may involve lengthier and more complex processes.

The time at which the differences between ERPs peak in stimulus-locked epochs, varies slightly from experiment to experiment, but tends to be between 500 and 750 ms after stimulus presentation. This is

[7] The response-locked epochs are base-line corrected at the response. So, here ERP amplitude is judged by comparing the minimum and the maximum voltage recorded.

slightly later than what was found in [6] but it is similar to the interval reported in [20]. The differences with the former study could be attributed to differences in tasks and stimuli.

The low p-values in the statistical test observed in the Pz grand averages and scalp maps, show that there are significant differences in the ERPs associated to different confidence levels that are common across experiments. For this reason such ERPs should represent the mental processes associated with a any (perceptual) decision making. The p-value scalp-maps in figure 4 show that ERPs associated with different confidence values start being statistically different at 550 ms after stimulus onset in the frontal area as found in [21]. Then, differences remain lateralised but spread to the parietal area, similar to what was reported in [23]. The lateralisation may be due to the fact that most participants in our experiments were right handed.

The p-value scalp-maps in figure 4 show that also the corresponding ERPs for response-locked epochs are lateralised, but also spread to central areas. These differences are most likely due to the high correlation between RT and confidence. For this reason, trials where decisions were made with higher confidence, have shorter RTs than trials where participants were less confident.

4.2. Confidence regression and transfer learning

In this study we were able to predict the confidence in a continuous way with better accuracy than the baseline (average reported confidence). This was true even for the approaches where the method was built with data from completely different subjects (MS), or different subjects and experiments (MT). We can see from figure 5 and from the statistical analyses performed, that, from the point of view of the MAE, the best approach to predict the confidence is, unsurprisingly, SS followed by MS and MT. However, from the point of view of the MCA and the calibration offset, the three approaches were not statistically different. For the $c\Delta$ measurement, the MT approach was inferior to the SS and MS approaches, but these were not statistically different from each other. This means that even if the accuracy of the prediction is better for the SS approach than the MS, their predictions have the same separability between correct and incorrect trials in terms of predicted confidence.

Furthermore, in figure 6 we can see that for both the SS and MS approaches, the ANN method predicts values across all the range of possibilities, with a distribution similar to the real one, a skewed distribution with many 'Sure' cases and an uniform-like distribution for the rest of the confidence levels. On the other hand, given that the baseline only predicts one value (the mean confidence) for each fold, the confusion matrices for the baseline show white spaces. For the MT approach, most of the ANN predictions are around the mean reported confidence, like the baseline, with a wider distribution similar to a Gaussian. Considering all these results, it appears that the MS approach is the most promising of the three. Even if it lacks the generalisation of MT, it is still a zero-training approach. This provides significant advantages towards creating 'plug-and-play' decision support systems, where the confidence of a participant could be predicted straight away. This could be used, for instance, to evaluate when the decision maker may need a break. Additionally, the confidence prediction from the SS and MS approach are equally good as surrogates of the accuracy.

Regarding the $c\Delta$ value obtained in different experiments (see bottom of figure 5), we observed a large standard deviation in every experiment, not only for the MT approach, but also for the ground truth. Interestingly, the predicted $c\Delta$ was not correlated with the real $c\Delta$ (p-value $= 0.255$), nor with the accuracy (p-value $= 0.051$). However, there was a correlation between the real $c\Delta$ value and the accuracy (p-value < 0.001). This suggest that the ability of the subject and, thus, the system to properly separate the correct and incorrect trials is directly impacted by the participant's accuracy.

Finally, considering the confusion matrices in figure 6, and the results when comparing the ANN method with the RT-Only, it appears that using only RT is sufficient for the SS approach, as it shows little variation within task and subject. However, it seems that the EEG helps to generalise the prediction which is needed for the MS and MT approaches. In particular, it can be seen in the confusion matrix for MT, that the prediction of the RT-Only method shows smaller variability on the prediction. For instance, the latter never predicts the 0.1 confidence level.

4.3. Regression models validation

Given that in both experiment SIMPLE-SEARCH-VALIDATION and REALISTIC-SEARCH-VALIDATION, the confidence was not reported, it is not possible to be certain of the quality of the prediction. Nevertheless, the meta-MCA of MT shows levels comparable to the ones obtained in the SS and MS approach, suggesting the accuracy might be similar to the one obtained with the eight training problems for which confidence was reported. This result supports the hypothesis that the confidence that can be detected through ERPs is not linked to the confidence-reporting process, but instead, to the actual decision-making process. This opens up the possibility of obtaining the users' confidence without having to ask for it directly after each decision. In addition, these results reinforce all the previous studies about neural correlates of confidence, suggesting that the features discovered in those studies were actually related to the decision-making process too. However, we cannot be certain whether the predicted confidence in the validation problems is equally useful as surrogate of the confidence like in the others

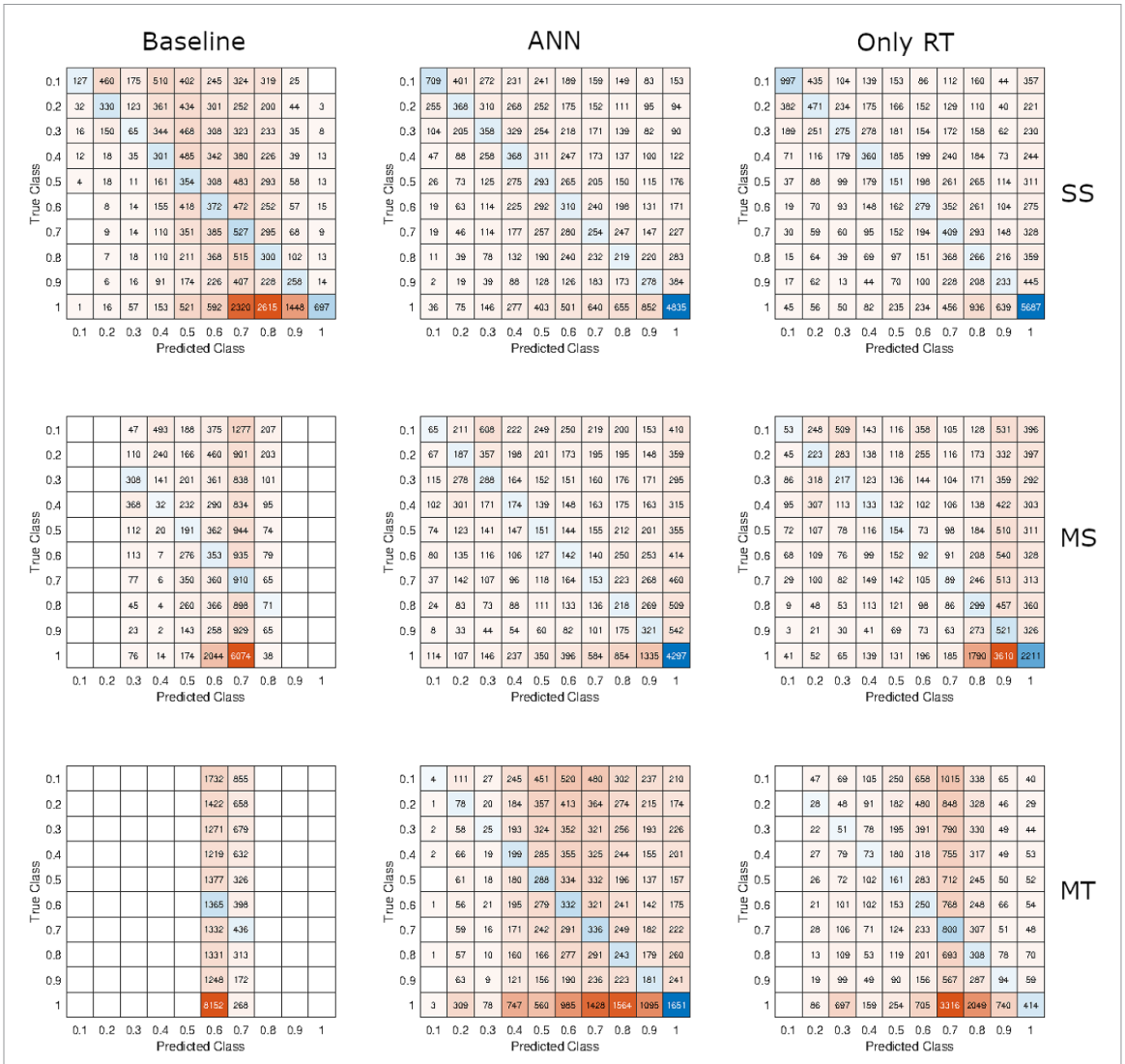

Figure 6. Confusion matrices for the three methods, Baseline, ANN, Only RT and three approaches, SS, MS and MT. The diagonal of each matrix correspond to the correctly predicted confidence values, while the other cells are incorrectly predicted confidences. The number in each cell indicates the number of trials falling in the corresponding class. The stronger the colour (red for incorrect and blue for correct) the more trials in a cell.

experiments. Even if the statistical analysis indicated that the $c\Delta$ was significantly larger than 0, if we look at the results of individual experiments (bottom of figure 5), we can see that these results vary significantly in different experiments (not only for the validation data, but for the training data as well). For example, SIMPLE-SEARCH-VALIDATION shows a $c\Delta$ similar to REALISTIC-SEARCH, OUTPOST, or ASSISTED-PATROL. However, REALISTIC-SEARCH-VALIDATION shows $c\Delta$ values that are not statistically different from 0. More experiments will be needed to find conclusive answers.

5. Conclusions

In this study we have shown that there is a difference in the ERPs elicited during the decision-making process for different confidence values. In particular, those trials where the participants answered that they were sure of their decision, showed the biggest differences. This points to the idea that there is more difference between being certain of a decision and having any degree of uncertainty, than the difference between different levels of uncertainty. This should be taken into consideration for the design of future experiments. Instead of allowing subjects a fine grain range of confidence levels, broader categories (one of them being 'Sure') show more separability in the EEG signals.

We have shown that some of the differences across confidence levels are present across experiments. Additionally, we have shown that it is possible to predict the confidence using only EEG signals and the RT, with better accuracy than the baseline. This is not only true for SS approaches, but also for zero training approaches such as predicting one participant's confidence using the rest of the participants as training data, or even more, predicting the confidence of every participant of a specific experiment

trained with other experiments. It is important to note that some of the experiments tested had notable differences in the stimulus presentation and one even had a different number of electrodes. Even if the MAE is higher for the zero-training approaches compare to the SS, the MS approach was not significantly different from the SS approach in terms of: MCA, $c\Delta$, and calibration offset. In the future, further investigation should be done on the limits of the transfer learning capabilities by, for example, testing more and more different experiments, or by removing part of the data. Also, the model should be improved in the future. In this study a shallow ANN was used. However deep networks have shown their potential on the field [62–64] making them a good candidate for future research on confidence prediction.

Finally, we further validated the model built by predicting the confidence in two experiments where the confidence was not reported. The predicted confidence had a $c\Delta$ significantly different from zero. This demonstrated that the predicted confidence was linked to the decision-making process rather than the confidence-reporting process. We realise that making a validation using a data from different experiments may not be always possible. However, we consider that this kind of validation is a step forward to have BCI systems used in less controlled environments and in real-world applications.

Acknowledgments

Disclaimer: This article is an overview of UK MOD sponsored research and is released for informational purposes only. The contents of this article should not be interpreted as representing the views of the UK MOD, nor should it be assumed that they reflect any current or future UK MOD policy. The information contained in this article cannot supersede any statutory or contractual requirements or liabilities and is offered without prejudice or commitment.

The authors acknowledge support of the UK Defence Science and Technology Laboratory (Dstl) and Engineering and Physical Research Council (EPSRC) under Grant No. EP/P009204/1. This is part of the collaboration between US DoD, UK MOD and UK EPSRC under the Multidisciplinary University Research Initiative. DV acknowledges support from the US DoD Bilateral Academic Research Initiative (BARI) under Grant No. W911NF1810434. We would also like to acknowledge and thank the contribution of Morgan Mason for the revision of this text.

ORCID iD

Jacobo Fernandez-Vargas https://orcid.org/0000-0002-6941-1233

References

[1] Gold J I and Shadlen M N 2007 The neural basis of decision making *Annu. Rev. Neurosci.* **30** 535–74
[2] de Lange F P, Jensen O and Dehaene S 2010 Accumulation of evidence during sequential decision making: the importance of top–down factors *J. Neurosci.* **30** 731–8
[3] Cheadle S, Wyart V, Tsetsos K, Myers N, De Gardelle V, Castañón S H and Summerfield C 2014 Adaptive gain control during human perceptual choice *Neuron* **81** 1429–41
[4] Pisauro M A, Fouragnan E, Retzler C and Philiastides M G 2017 Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous EEG-fMRI *Nat. Commun.* **8** 15808
[5] Polanía R, Krajbich I, Grueschow M and Ruff C C 2014 Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making *Neuron* **82** 709–20
[6] Graziano M, Parra L C and Sigman M 2015 Neural correlates of perceived confidence in a partial report paradigm *J. Cogn. Neurosci.* **27** 1090–103
[7] Krumpe T, Gerjets P, Rosenstiel W and Spüler M 2020 Decision confidence: EEG correlates of confidence in different phases of an old/new recognition task *Brain-Computer Interfaces* **6** 162–77
[8] Kepecs A, Uchida N, Zariwala H A and Mainen Z F 2008 Neural correlates, computation and behavioural impact of decision confidence *Nature* **455** 227–31
[9] Adler W T and Ma W J 2018 Comparing Bayesian and non-Bayesian accounts of human confidencereports *PLoS Comput. Biol.* **14** 1–34
[10] Sanders J I, Hangya B and Kepecs A 2016 Signatures of a statistical computation in the human sense of confidence *Neuron* **90** 499–506
[11] Faivre N, Filevich E, Solovey G, Kühn S and Blanke O 2018 Behavioral, modeling, and electrophysiological evidence for supramodality in human metacognitio *J. Neurosci.* **38** 263–77
[12] Grimaldi P, Lau H and Basso M A 2015 There are things that we know that we know, and there are things that we do not know we do not know: confidence in decision-making *Neurosci. Biobehavioral Rev.* **55** 88–97
[13] Ratcliff R and Starns J J 2013 Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination *Psychol. Rev.* **120** 697–719
[14] Baranski J V and Petrusic W M 1994 The calibration and resolution of confidence in perceptual judgments *Percept. Psychophys.* **55** 412–28
[15] Aitchison L, Bang D, Bahrami B and Latham P E 2015 Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making *PLoS Comput. Biol.* **11** 1–23
[16] Christopoulos G I, Tobler P N, Bossaerts P, Dolan R J and Schultz W 2009 Neural correlates of value, risk, and risk aversion contributing to decision making under risk *J. Neurosci.* **29** 12574–83
[17] Philiastides M G and Sajda P 2006 Temporal characterization of the neural correlates of perceptual decision making in the human brain} *Cereb. Cortex* **16** 509–18
[18] Boldt A and Yeung N 2015 Shared neural markers of decision confidence and error detection *J. Neurosci.* **35** 3478–84
[19] Molenberghs P, Trautwein F M, Böckler A, Singer T and Kanske P 2016 Neural correlates of metacognitive ability and of feeling confident: a large-scale fMRI study *Soc. Cogn. Affect. Neurosci.* **11** 1942–51
[20] Gherman S and Philiastides M G 2015 Neural representations of confidence emerge from the process of decision formation during perceptual choices *Neuroimage* **106** 134–43
[21] Gherman S and Philiastides M G 2018 Human VMPFC encodes early signatures of confidence in perceptual decisions *eLife* **7** e38293

[22] Basten U, Biele G, Heekeren H R and Fiebach C J 2010 How the brain integrates costs and benefits during decision making *Proc. Natl Acad. Sci.* **107** 21767–72
[23] Herding J, Ludwig S, von Lautz A, Spitzer B and Blankenburg F 2019 Centro-parietal EEG potentials index subjective evidence and confidence during perceptual decision making *Neuroimage* **201** 116011
[24] Kiani R and Shadlen M N 2009 Representation of confidence associated with a decision by neurons in the parietal cortex *Science* **324** 759–64
[25] Wang X W, Nie D and Lu B L 2014 Emotional state classification from EEG data using machine learning approach *Neurocomputing* **129** 94–106
[26] Berka C, Levendowski D J, Lumicao M N, Yau A, Davis G, Zivkovic V T, Olmstead R E, Tremoulet P D and Craven P L 2007 EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks *Aviat. Space Environ. Med.* **78** B231–44
[27] Tremmel C, Herff C, Sato T, Rechowicz K, Yamani Y and Krusienski D J 2019 Estimating Cognitive Workload in an Interactive Virtual Reality Environment Using EEG *Front. Hum. Neurosci.* **13** 401
[28] Ojanen V, Revonsuo A and Sams M 2003 Visual awareness of low-contrast stimuli is reflected in event-related brain potentials *Psychophysiology* **40** 192–7
[29] Yeung N and Summerfield C 2012 Shared neural markers of decision confidence and error detection *Phil. Trans. R. Soc.* B **367** 1310–21
[30] Vi C and Subramanian S 2012 Detecting error-related negativity for interaction design *Proc. Conf. Human Factors in Computing Systems* (*Austin, TX, May 2012*) pp 493–502
[31] Kubanek J, Hill J, Snyder L H and Schalk G 2015 Cortical alpha activity predicts the confidence in an impending action *Front. Neurosci.* **9** 243
[32] Wolpaw J R, Birbaumer N, McFarland D J, Pfurtscheller G and Vaughan T M 2002 Brain-computer interfaces for communication and control *Clin. Neurophysiol.* **113** 767–91
[33] Müller-Putz G, Schwarz A, Pereira J and Ofner P 2016 From classic motor imagery to complex movement intention decoding: the noninvasive Graz-BCI approach *Progress in Brain Research* (Amsterdam: Elsevier) pp 1–32
[34] Birbaumer N 2006 Breaking the silence: Brain-computer interfaces (BCI) for communication and motor control *Psychophysiology* **43** 517–32
[35] Vaid S, Singh P and Kaur C 2015 EEG signal analysis for BCI interface: a review *Int. Conf. Advanced Computing and Communication Technologies, ACCT* vol 2015-April (Institute of Electrical and Electronics Engineers Inc.) pp 143–7
[36] Cinel C, Valeriani D and Poli R 2019 Neurotechnologies for human cognitive augmentation: currentstate of the art and future prospects *Front. Hum. Neurosci.* **13** 13
[37] Poli R, Valeriani D and Cinel C 2014 Collaborative brain-computer interface for aiding decision-making *PLoS One* **9** 7
[38] Valeriani D, Poli R and Cinel C 2017 Enhancement of group perception via a collaborativebrain–computer interface *IEEE Trans. Biomed. Eng.* **64** 1238–48
[39] Valeriani D, Cinel C and Poli R 2017 Group augmentation in realistic visual-search decisions viaa hybrid brain-computer interface *Sci. Rep.* **7** 7772
[40] Valeriani D and Poli R 2019 Cyborg groups enhance face recognition in crowdedenvironments *PLoS One* **14** 1–17
[41] Valeriani D and Matran-Fernandez A 2018 Past and future of multi-mind brain-computer interfaces *Brain-Computer Interfaces Handbook: Technological and Theoretical Advances*, ed C S Nam, A Nijholt and F Lotte (Boca Raton, FL: CRC Press) ch 36, pp 685–700
[42] Pan S J and Yang Q 2009 A survey on transfer learning *IEEE Trans. Knowl. Data Eng.* **22** 1345–59
[43] Waytowich N R, Lawhern V J, Bohannon A W, Ball K R and Lance B J 2016 Spectral transferlearning using information geometry for a user-independent brain-computer interface *Front. Neurosci.* **10** 430
[44] Ming Y, Ding W, Pelusi D, Wu D, Wang Y K, Prasad M and Lin C T 2019 Subject adaptation network for EEG data analysis *Appl. Soft Comput.* **84** 105689
[45] Dai M, Zheng D, Liu S and Zhang P 2018 Transfer kernel common spatial patterns for motorimagery brain-computer interface classification *Comput. Math. Methods Med.* **2018** 9871603
[46] Jayaram V, Alamgir M, Altun Y, Scholkopf B and Grosse-Wentrup M 2016 Transfer learningin brain-computer interfaces *IEEE Comput. Intell. Mag.* **11** 20–31
[47] Rodrigues P L C, Jutten C and Congedo M 2018 Riemannian procrustes analysis: Transfer learningfor brain–computer interfaces *IEEE Trans. Biomed. Eng.* **66** 2390–401
[48] Rolls E T, Grabenhorst F and Deco G 2010 Choice, difficulty, and confidence in the brain *Neuroimage* **53** 694–706
[49] Valeriani D, Poli R and Cinel C 2015 A collaborative brain-computer interface for improving group detection of visual targets in complex natural environments *2015 7th Int. IEEE/Conf. Neural Engineering (NER)* (IEEE) pp 25–28
[50] Bhattacharyya S, Valeriani D, Cinel C, Citi L and Poli R 2019 Target detection in video feeds with selected dyads and groups assisted by collaborative brain-computer interfaces *2019 9th Int. IEEE/Conf. Neural Engineering (NER)* pp 159–62
[51] Bhattacharyya S, Valeriani D, Cinel C, Citi L and Poli R 2019 Collaborative brain-computer interfaces to enhance group decisions in an outpost surveillance task *2019 41st Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)* pp 3099–102
[52] Cui Y, Xu Y and Wu D 2019 EEG-based driver drowsiness estimation using feature weightedepisodic training *IEEE Trans. Neural Syst. Rehabil. Eng.* **27** 2263–73
[53] Wu D, Lance B J, Lawhern V J, Gordon S, Jung T and Lin C 2017 EEG-based user reactiontime estimation using riemannian geometry features *IEEE Trans. Neural Syst. Rehabil. Eng.* **25** 2157–68
[54] Fernandez-Vargas J, Kita K and Yu W 2016 Real-time hand motion reconstruction system fortrans-humeral amputees using EEG and EMG *Front. Robot. AI* **3** 50
[55] Jervis B W, Ifeachor E C and Allen E M 1988 The removal of ocular artefacts from theelectroencephalogram: a review *Med. Biol. Eng. Comput.* **26** 2–12
[56] Luce R D R D 1986 *Response Times: Their Role in Inferring Elementary Mental Organization* (Oxford: Oxford University Press)
[57] Kiani R, Corthell L and Shadlen M 2014 Choice certainty is informed by both evidence and decisiontime *Neuron* **84** 1329–42
[58] Fernandez-Vargas J, Tarvainen T V J, Kita K and Yu W 2017 Effects of UsingVirtual Reality and Virtual Avatar on Hand Motion Reconstruction Accuracy andBrain Activity *IEEE Access* **5** 23736–50
[59] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: A simpleway to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58
[60] Fleming S M, Huijgen J and Dolan R J 2012 Prefrontal contributions to metacognition inperceptual decision making *J. Neurosci.* **32** 6117–25
[61] Fleming S M and Lau H C 2014 How to measure metacognition, *Front. Hum. Neurosci.* **8** 443
[62] Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A and Yger F 2018 A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update *J. Neural Eng.* **15** 031005
[63] Roy Y, Banville H, Albuquerque I, Gramfort A, Falk T H and Faubert J 2019 Deep learning-based electroencephalography analysis: a systematic review *J. Neural Eng.* **16** 051001
[64] Craik A, He Y and Contreras-Vidal J L 2019 Deep learning for electroencephalogram (EEG) classification tasks: a review *J. Neural Eng.* **16** 031001