PAPER

Multimodal collaborative brain-computer interfaces aid human-machine team decision-making in a pandemic scenario

View the article online for updates and enhancements.

Journal of Neural Engineering

PAPER

Multimodal collaborative brain-computer interfaces aid human-machine team decision-making in a pandemic scenario

Davide Valeriani[1] , Lena C O'Flynn[1,3], Alexis Worthley[1], Azadeh Hamzehei Sichani[1] and Kristina Simonyan[1,2,3,*]

[1] Department of Otolaryngology-Head and Neck Surgery, Massachusetts Eye and Ear and Harvard Medical School, 243 Charles Street, Boston MA 02114, United States of America
[2] Department of Neurology, Massachusetts General Hospital, 55 Fruit Street, Boston MA 02114, United States of America
[3] Program in Speech Hearing Bioscience and Technology, Harvard University, 260 Longwood Avenue, Boston MA 02115, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: kristina_simonyan@meei.harvard.edu

Abstract

Objective. Critical decisions are made by effective teams that are characterized by individuals who trust each other and know how to best integrate their opinions. Here, we introduce a multimodal brain-computer interface (BCI) to help collaborative teams of humans and an artificial agent achieve more accurate decisions in assessing danger zones during a pandemic scenario. *Approach.* Using high-resolution simultaneous electroencephalography/functional MRI (EEG/fMRI), we first disentangled the neural markers of decision-making confidence and trust and then employed machine-learning to decode these neural signatures for BCI-augmented team decision-making. We assessed the benefits of BCI on the team's decision-making process compared to the performance of teams of different sizes using the standard majority or weighing individual decisions. *Main results.* We showed that BCI-assisted teams are significantly more accurate in their decisions than traditional teams, as the BCI is capable of capturing distinct neural correlates of confidence on a trial-by-trial basis. Accuracy and subjective confidence in the context of collaborative BCI engaged parallel, spatially distributed, and temporally distinct neural circuits, with the former being focused on incorporating perceptual information processing and the latter involving action planning and executive operations during decision making. Among these, the superior parietal lobule emerged as a pivotal region that flexibly modulated its activity and engaged premotor, prefrontal, visual, and subcortical areas for shared spatial-temporal control of confidence and trust during decision-making. *Significance.* Multimodal, collaborative BCIs that assist human-artificial agent teams may be utilized in critical settings for augmented and optimized decision-making strategies.

1. Introduction

Every human decision is accompanied by a degree of confidence, which represents how likely that decision is to be correct (Fleming and Lau 2014, Pouget *et al* 2016). The ability to compute such likelihood is, however, very variable between individuals. It is also subject to various biases as some individuals take into account the uncertainty of their perception (Navajas *et al* 2017) while others tend to underestimate or overestimate their

confidence (Moore and Cain 2007). To minimize implicit variability and bias, critical decisions are often made by teams that aggregate individual opinions based on the confidence of each member using a weighted majority voting (Surowiecki 2004, Meyen *et al* 2021). Moreover, effective collaborations rely on information sharing and mutual trust between the team members, attributing to the individual willingness to reciprocate the decisions of others (Krueger *et al* 2007, Park and Lee 2014, Pescetelli and Yeung 2021).

In examining neural representations of a decision-making process, recent neuroimaging studies have shown that decision confidence is encoded in the midline parietal, prefrontal, and insular cortices (Boldt and Yeung 2015, Bang and Fleming 2018, Gherman and Philiastides 2018, Shekhar and Rahnev 2018, Pereira *et al* 2020), partly overlapping with neural correlates of first-order decisions and change of mind (Fleming *et al* 2018, Pereira *et al* 2020). On the other hand, neural representations of trust appear to be linked to the activity of prefrontal, paracingulate and insular cortices, the septal area, and ventral tegmental area (Krueger *et al* 2007, van den Bos *et al* 2009), intersecting with those of decision confidence.

In recent years, machine learning-based decoders of decision confidence (Krumpe *et al* 2020, Fernandez-Vargas *et al* 2021) and brain-computer interfaces (BCIs) for improving decision-making accuracy (Valeriani *et al* 2017, Valeriani and Poli 2019) have increasingly utilized the neural signatures of decision-making and trust. However, the existing technologies have been limited to decoding individual neural activity without considering the team's metacognitive capabilities due, in part, to an incomplete understanding of neural signatures of decision-making strategies and trust among the team members.

Given the potential of collaborative BCIs (e.g. human–human, human–artificial agent) for augmenting individual performance in critical decision-making, we developed a BCI to improve team decision for the identification of danger zones needing resource allocation in a realistic pandemic scenario and compared its benefits with those provided by an artificial agent. Specifically, we used simultaneous electroencephalography (EEG)/functional MRI (fMRI) in healthy individuals to investigate the behavioral and spatial-temporal neural underpinnings of a team's perceptual decision-making in the pandemic scenario. We examined the accuracy and subjective confidence of the team with and without feedback from the artificial agent, as well as its trust within the compliance (Drnec *et al* 2016) and agreeableness (Evans and Revelle 2008) frameworks. We hypothesized that to improve team decisions in critical scenarios, the BCI should accurately estimate the accuracy of each team member (BCI confidence) and cross-monitor trust formation by leveraging neural markers of team decision-making. To test this hypothesis, we evaluated the team BCI using the identified spatio-temporal neural patterns of decoded decision confidence and trust to augment team decision-making.

2. Material and methods

2.1. Participants
Fourteen healthy volunteers (7 females/7 males, mean age 42.7 ± 12.6 yr) participated in this study. All participants were native English speakers, right-handed, and had no past or present history of neurological or psychiatric disorders. Informed, written consent was obtained before study participation according to the procedures approved by the Institutional Review Board of Mass General Brigham.

2.2. Experimental design
Each participant underwent a perceptual decision-making experiment composed of six blocks of 30 trials. Each trial started with a fixation cross, displayed for 500 ms, followed by a fictional geographical map of two regions (*empty map*) for another 500 ms, during which the participants familiarized themselves with the location and area of these two regions (figure 1(A)). The displayed empty map was chosen from a library of 40 maps created for this study using the publicly available Fantasy Map Generator online tool (https://azgaar.github.io/Fantasy-Map-Generator/) and custom Python 3.6 software. In the following 500 ms, dots representing pandemic cases were presented on the map (*stimulus*), with the dot color (in RGB format, with values from 0 to 255) indicating the pandemic severity and ranging from mild (white, RGB = 255, 255, 255) to critical (red, RGB = 255, 0, 0) on a continuous scale. The number of dots varied between 15 and 40 per trial (mean 25.7 ± 7.3 dots). The difficulty of each trial was controlled by changing the difference in the number of dots in each region (range 0–14) and the difference in the severity of each region, represented by the difference between 255 and the green or blue component of the RGB color value of each dot (range 103–998). After the stimulus presentation, participants were asked to decide as quickly as possible within 2 s which region (1 or 2) was most in danger and to log their decision (*first decision*) using a two-button keypad. Participants were then asked to report within 2 s their degree of confidence in that decision (*confidence*), ranging from 1 (not confident) to 4 (very confident), using a four-button keypad. Trials with each region (1 or 2) being in the most danger represented 50% of the trials of each block. After a jittered delay of 2.5–3.0 s (*jitter*), a feedback screen was presented for 2 s displaying participants' decision and confidence compared to the decision and confidence of a pre-constructed artificial agent (*feedback*). The artificial agent linearly increased its accuracy over the trial blocks, from 56% accuracy in the first block to 90% accuracy in the last block, to mimic the performance of a well-trained human while promoting trust formation. The artificial agent was highly calibrated when estimating its own confidence. When correct, the agent provided confidence with the ratings of 4 (very confident) in 40% of trials, 3 (confident) in 30% of trials, 2 (somewhat confident) in 20% of trials, and 1 (not confident) in 10% of trials. When incorrect, the agent provided confidence with the ratings of 4 (very confident) in 10% of trials, 3 (confident) in 20% of trials, 2 (somewhat confident) in 30% of trials,
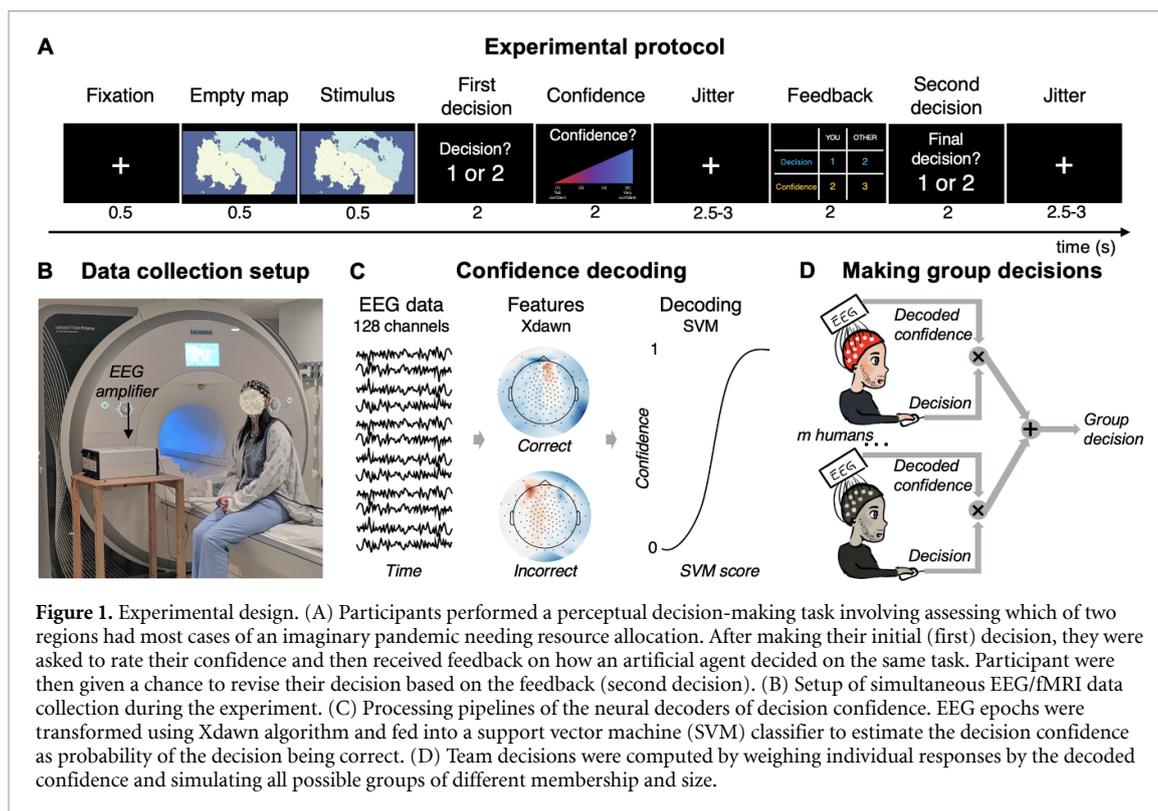
Figure 1. Experimental design. (A) Participants performed a perceptual decision-making task involving assessing which of two regions had most cases of an imaginary pandemic needing resource allocation. After making their initial (first) decision, they were asked to rate their confidence and then received feedback on how an artificial agent decided on the same task. Participant were then given a chance to revise their decision based on the feedback (second decision). (B) Setup of simultaneous EEG/fMRI data collection during the experiment. (C) Processing pipelines of the neural decoders of decision confidence. EEG epochs were transformed using Xdawn algorithm and fed into a support vector machine (SVM) classifier to estimate the decision confidence as probability of the decision being correct. (D) Team decisions were computed by weighing individual responses by the decoded confidence and simulating all possible groups of different membership and size.

and 1 (not confident) in 40% of trials. This resulted in a type-2 area under the receiver operating characteristic (ROC) curve (AUC) of 67.3% for the artificial agent, demonstrating its high calibration. Based on this feedback, participants were asked to reconsider their decision about which region was most in danger within 2 s (*second decision*) and then rest for a jittered period of 2.5–3.0 s (*jitter*). At the end of each block, participants were also asked to evaluate their trust in the artificial agent using a four-button keypad, with the responses ranging from 1 (low trust) to 4 (high trust).

Prior to the experiment, participants were familiarized with the experimental tasks by completing one training block of 10 trials that included maps that were similar to those used during the actual experiment. To avoid bias in the participants' behavior during the experiment, participants were not familiarized with the artificial agent's feedback during training.

2.3. Data acquisition

Six simultaneous EEG/fMRI blocks were acquired in each participant using an MR-compatible 128-electrode EEG system (Magstim, Inc) in a 3.0 Tesla Siemens Prisma scanner equipped with a 20-channel head coil (figure 1(B)). The experimental paradigm (figure 1(A)) was presented using E-Prime 2.0 software with the Net Station 2.0 extension and the Hyperion MRI digital projection system with millisecond accuracy (Psychology Software Tools, Sharpsburg, PA). Head movements during scanning were minimized by tightly cushioning and restricting the participant's head inside the coil and instructing the participant to minimize the head and body motions throughout the scan.

EEG data were obtained at a sampling rate of 1000 Hz using NetStation 5.4.2. Impedances were kept below 100 kΩ by soaking the sponge electrodes in a saline solution before fitting onto the participant's head (Ferree *et al* 2001). Each electrode was visually inspected to avoid loops before placing the participant in the MRI scanner. EEG data were synchronized to the MRI scanner's 10 MHz clock, and the onset time of every fMRI volume was marked in the EEG acquisition. The electrocardiographic signal was synchronized with EEG and recorded using two MR-compatible electrodes. Eye saccades were minimized by instructing participants to focus their gaze on the center of the presentation screen. All EEG epochs were visually inspected offline to ensure their quality, which resulted in discarding on average 4.4% ± 2.8% stimulus-locked epochs and, on average, 1.1% ± 2.4% feedback-locked epochs.

fMRI data were acquired using a gradient-weighted echo-planar imaging (EPI) pulse sequence (70 slices, voxel size 1.7 × 1.7 × 2.5 mm, repetition time (TR) 2,200 ms, echo time (TE) 26 ms, flip angle (FA) 80°, field-of-view (FOV) 220 × 220 mm). A high-resolution whole-brain T1-weighted structural image was collected using a three-dimensional magnetization-prepared rapid acquisition with gradient echo sequence (224 slices, voxel size 0.8 mm isotropic, TR 2,400 ms, TE 2.07 ms, FA 8°, FOV 179 × 240 mm) as an fMRI/EEG reference.

Participants' responses throughout the experiment were recorded to evaluate behavioral correlates of decision-making accuracy (i.e. the averaged probability of correct responses), confidence, and trust.

2.4. Data preprocessing

Gradient and ballistocardiogram artifacts were removed from EEG data using template subtraction (Allen *et al* 2000) and optimal basis sets (Niazy *et al* 2005) methods implemented in NetStation 5.4.2 software. EEG data were referenced to Cz and band-pass filtered between 1 and 40 Hz using a zero-phase, non-causal filter with a length of 3301 samples implemented in the MNE 0.19.1 Python package (Gramfort *et al* 2013). EEG data from each trial were segmented into two types of epochs: stimulus-locked and feedback-locked, each lasting 1.5 s. Stimulus-locked epochs started with the onset of the *empty map* stimulus, while feedback-locked epochs started at the onset of the *feedback* screen (figure 1(A)). Each epoch was corrected to baseline by subtracting the average voltage recorded in a 100 ms interval preceding the epoch's onset and downsampled to 100 Hz. Epochs with peak-to-peak amplitude bigger than 5 mV were considered contaminated by motion artifacts and therefore discarded (Pisauro *et al* 2017). Epochs associated with the trials where the participants did not report their *first decision* (7.0% ± 4.5%) were discarded.

The fMRI data were analyzed using the *afni_proc.py* processing pipeline of AFNI software. In brief, the first two volumes of each block were removed to account for the magnetization equilibrium. Spikes in time series were truncated, and the remaining data were registered to the volume collected closest in time to the anatomical scan using heptic polynomial interpolation. Time series were then aligned to the skull-stripped anatomical image, spatially normalized to the AFNI standard Talairach-Tournoux space, spatially smoothed with a 4 mm Gaussian filter, and normalized to the percent signal change. To control for motion artifacts, six motion parameters calculated during the realignment of the EPI volumes were used in the regression model, TRs with Euclidean norm of the derivative of the motion parameters exceeding 1.0 were excluded, and TRs with more than 3% of the voxels marked as outliers were censored. As a result of this three-step motion correction, one subject with more than 40% of censored TRs was excluded from the final fMRI analysis. As the last step, in each subject, a separate linear regressor for each stimulus (i.e. accuracy, confidence, trust) was convolved with a canonical hemodynamic response function for confidence and trust to account for different stimuli durations and to derive the neural response.

2.5. Statistical analysis

2.5.1. EEG data

For each participant and EEG channel, we computed the median EEG epoch for trials where participants (a) made the correct decision, (b) made the incorrect decision, (c) reported a low level of confidence (1 or 2), (d) reported a high level of confidence (3 or 4), (e) trusted the agent, (f) distrusted the agent, by averaging the voltages recorded at each time point in stimulus-locked (a)–(d) or feedback-locked (e)–(f) epochs. Grand median signals were computed for each EEG channel across participants. Statistical analysis was conducted by comparing the participant's median signals between the different conditions (a–b, c–d, e–f) using the Wilcoxon signed-rank test at $p \leqslant 0.05$. For each stimulus- and feedback-locked epochs, we computed the time-frequency power spectral density using the multitaper method with a time bandwidth of 4 s. We assessed significant differences between correct-incorrect, confident-not confident, and trust-distrust spectrograms using the Wilcoxon signed-rank test at $p \leqslant 0.05$. Further, for each epoch we computed the power in the delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz) frequency bands normalized by the total power in the range of 1–30 Hz. Statistical analysis between the median power in each frequency band was conducted using the Wilcoxon signed-rank test at $p \leqslant 0.0125$ to correct for multiple comparisons (0.05/4 power bands).

2.5.2. fMRI data

To identify neural markers of accuracy, we contrasted trials in which participants made a correct decision vs. trials with an incorrect decision and performed a paired two-tailed *t*-test at corrected $p \leqslant 0.05$, with a voxelwise threshold of $p \leqslant 0.01$ and a minimum cluster size $\geqslant 168$ mm^3. To identify neural markers of subjective confidence, we contrasted trials in which participants reported a low level of confidence (values 1 and 2) vs. a high level of confidence (values 3 and 4) and performed a paired two-tailed *t*-test at corrected $p \leqslant 0.05$, with voxelwise threshold $p \leqslant 0.01$, minimum cluster size $\geqslant 1194$ mm^3. Finally, neural markers of trust were assessed by contrasting trials where participants expressed trust vs. distrust in the agent using a paired two-tailed *t*-test at corrected $p \leqslant 0.05$, with voxelwise threshold $p \leqslant 0.01$ and a minimum cluster size $\geqslant 168$ mm^3.

2.5.3. Behavioral data

Accuracy was distinguished between the first (pre-feedback) and second (post-feedback) decisions. We performed a two-sided Wilcoxon signed-rank test at $p \leqslant 0.05$ to determine if one response was more

likely to be chosen than another (decision bias). Accuracy (%) was computed by comparing each individual decision to the correct answer and then averaging across trials and participants. We performed a two-sided Wilcoxon signed-rank test at $p \leqslant 0.05$ to determine the significant difference between the decision accuracy levels and, therefore, measure whether the feedback had a significant effect on the participant's decisions. Additionally, accuracy was computed separately per block and analyzed as a measure of task learning, i.e. accuracy should increase over blocks as participants get more familiar with the task. We tested the significance of such an effect with Friedman's chi-square test at $p \leqslant 0.05$. To compare participants' response times (in ms) during correct and incorrect trials, we conducted a two-sided Wilcoxon signed-rank test at $p \leqslant 0.05$. We further used Friedman's chi-square test to determine whether subjective confidence was modulated by accuracy, as well as whether subjective confidence was modulated by response time at $p \leqslant 0.05$. The degree of association between accuracy and subjective confidence was examined using the type-2 AUC (Fleming and Lau 2014). Finally, we conducted a Friedman's chi square test to assess whether trust significantly increased over blocks at $p \leqslant 0.05$.

2.6. BCI setup

2.6.1. Confidence decoding
Accuracy was decoded from EEG stimulus-locked epochs using a BCI composed of two modules: neural feature extraction and confidence estimation (figure 1(C)). In each participant, we used the epochs extracted from the first three blocks of the experiment as a training set and the epochs from the last three blocks as a testing set (temporal splitting) to simulate a realistic (online) scenario where a BCI would undergo a training phase before starting to use the system. Each epoch was labeled as 'confident' or 'not confident', corresponding to the participant's decision during the experiment. To extract the neural features from the stimulus-locked epochs, we used the Xdawn spatial filtering method (Rivet *et al* 2009) as implemented in the MNE 0.19.1 Python package (Gramfort *et al* 2013). Xdawn decomposed the 128-channel epochs into 16-component epochs that maximized the separation between the 'confident' and 'not confident' classes. The number of components was chosen using a grid-search strategy (range 4–20) to maximize the decoding accuracy of the training set. The components of each epoch were then concatenated over the time domain to form the feature vector of that trial. A support vector machine (SVM) with radial basis function kernel and regularization parameter $C = 1000$ was used to transform each participant's features into a probability of a 'confident' trial. Specifically, the SVM was trained to predict from the Xdawn features whether the decision

made by the participant was correct or incorrect. We then used the Platt scaling method to compute the probability associated with the prediction of the correct class, which we interpreted as BCI confidence of that decision (Platt 1999). Xdawn and SVM were trained on data from the training set and applied to the test dataset to estimate the BCI confidence of each participant in the test trials. SVM was chosen among other classifiers as it is broadly used in real-time BCIs and typically outperforms other classifiers (Lotte *et al* 2018).

2.6.2. Group decision making
For each group size of $m = 2–14$, we assembled all possible teams that could be formed with 14 participants. Team decisions were formulated as follows:

$$d_{\text{group}} = \text{sign}\left(\sum_{i=1}^{m} d_i w_i\right),$$

where d_i is the decision of group member i, assuming value -1 for region 1 and $+1$ for region 2, and w_i is the weight associated with that decision. We compared team decisions made using the majority, confidence, and BCI methods (figure 1(D)) for all possible groups we could assemble with our $N = 14$ participants, that is, $\begin{pmatrix} N \\ k \end{pmatrix}$ for groups of size k. The majority had $w_i = 1$, $\forall i$, hence assigning the same weight to all participants. The confidence had weights equal to the confidence reported by participants after each decision. The BCI used the decoded confidence probabilities estimated by the SVM from EEG neural features as weights for each decision. For each team of a given size, the accuracy in the test set was computed using the three different methods described above to establish the weights and the first/second response of the participant as d_i. Because the participants' report on confidence was not collected after the second response, the confidence method was only applied to the first decision. For each team size 2–14, we computed the average accuracy across groups of that size in the test set using the five different decision-making strategies (i.e. majority first decision, confidence first decision, BCI first decision, majority second decision, and BCI second decision).

2.6.3. Trust estimation
We divided trials into two groups depending on their influence on the participant's trust in the artificial agent. Trust trials were those in which both the participant and the agent made either the same first and second decisions (agreeableness) or opposite first decision, followed by the change of mind of the participant during the second decision (compliance) (figure 5(A)). Distrust trials were those where the participant's final decision was different from the agent's decision (figure 5(A)). We assigned the value of 1 to

distrust trials and the value of 4 to trust trials. We then averaged the trust values using a rolling window of 5 trials to construct a behavioral model of trust for each participant. Averages of model-based trust within each block were compared with the trust reported by each participant at the end of that block using Spearman's correlation at $p \leqslant 0.05$. We also assessed whether the trust estimated by the behavioral model and the trust reported by the participants increased over time due to a familiarization with the agent's behavior using the Friedman's test at $p \leqslant 0.008$, corrected for multiple comparisons (0.05/6 observations of reported trust).

3. Results

3.1. Behavioral performance

Participants achieved an average accuracy of 61.9% in their first decision related to the geographic distribution of the pandemic severity compared to the artificial agent that was programmed to have an average accuracy of 73.3% (figure 2(A)). Following the feedback from the agent, participants' average accuracy improved slightly but not significantly to 65.2% ($W = 18.5$, Cohen's $d = -0.31$, $p = 0.06$) over the duration of six experimental blocks (figure 2(B)). Participants' decisions were not biased towards region 1 or region 2: the average number of responses for region 1 was 88.9 and for region 2 was 85.9 ($W = 45.0$, Cohen's $d = 0.19$, $p = 1.0$).

Response times did not correlate with accuracy of the first or second decisions (figure 2(C), first decision $W = 49.0$, Cohen's $d = -0.01$, $p = 0.86$; second decision $W = 35.0$, Cohen's $d = 0.23$, $p = 0.30$). Participants' confidence varied depending on their accuracy ($W = 0.58$, $Q = 20.7$, $p = 0.0001$) (figure 2(D)). The type-2 AUC of each participant was significantly better than random (0.58 ± 0.07, $T = 4.2$, Cohen's $d = 1.13$, $p = 0.001$). Response times were modulated by reported confidence ($W = 0.41$, $Q = 14.8$, $p = 0.002$) (figure 2(C)). There were no significant differences in response times ($W = 43.0$, Cohen's $d = 0.07$, $p = 0.58$) or reported confidence ($W = 30.0$, Cohen's $d = -0.20$, $p = 0.17$) between trust and distrust trials. The trust in the artificial agent reported by participants showed no significant change over time ($W = 0.05$, $Q = 3.6$, $p = 0.61$) (figure 2(E)).

As expected from the Condorcet theorem, the average team accuracy increased with the team size for all decision strategies (figure 3(A)). Weighing the first responses of individuals using reported confidence showed an improvement in accuracy over the standard majority, which was statistically significant for all team sizes that included 2 participants ($W = 528.5$, Cohen's $d = -0.32$, $p = 6.0 \times 10^{-10}$), 4 participants

($W = 85\,507.5$, Cohen's $d = -0.31$, $p = 6.3 \times 10^{-73}$), 5 participants ($W = 818\,589.0$, Cohen's $d = -0.09$, $p = 1.2 \times 10^{-12}$), 6 participants ($W = 995\,309.5$, Cohen's $d = -0.30$, $p = 5.8 \times 10^{-155}$), 7 participants ($W = 2\,344\,056.0$, Cohen's $d = -0.12$, $p = 3.7 \times 10^{-25}$), 8 participants ($W = 1\,172\,633.0$, Cohen's $d = -0.29$, $p = 6.0 \times 10^{-115}$), 9 participants ($W = 743\,690.5$, Cohen's $d = -0.18$, $p = 1.4 \times 10^{-23}$), 10 participants ($W = 131\,190.5$, Cohen's $d = -0.33$, $p = 4.8 \times 10^{-39}$), 11 participants ($W = 23\,470.5$, Cohen's $d = -0.22$, $p = 1.2 \times 10^{-6}$), or 12 participants ($W = 915.5$, Cohen's $d = -0.46$, $p = 3.2 \times 10^{-6}$). Weighing individual decisions using BCI confidence estimates significantly increased the accuracy of all team sizes of 2–13 participants with respect to both reported confidence (all $W > 0.0$, Cohen's $d < -0.21$, $p < 0.006$) and standard majority (all $W > 0.0$, Cohen's $d < -0.36$, $p < 1.3 \times 10^{-4}$). The magnitude of improvement further increased with the team size when comparing BCI vs. majority decision ($R_S = 0.92$, Cohen's $d = -2.52$, $p = 4.4 \times 10^{-6}$) and BCI vs. confidence ($R_S = 0.96$, Cohen's $d = -2.52$, $p = 5.1 \times 10^{-8}$), suggesting that larger benefits were provided by the BCI in larger groups.

Team decisions based on the second response from each participant were significantly more accurate than team decisions based on the first response for both the BCI and the majority methods for all group sizes of 2–13 participants (all $W > 0.0$, Cohen's $d < -0.46$, $p < 1.2 \times 10^{-4}$). However, individual accuracies did not differ significantly ($W = 18.5$, Cohen's $d = -0.31$, $p = 0.06$) (figure 2(A)). Pairs of participants who used the first decision (no artificial agent feedback) and the BCI confidence (figure 3(A)) were not significantly different from those pairs of participants who used the second decision (after artificial agent feedback) and standard majority ($W = 2061.0$, Cohen's $d = 0.04$, $p = 0.90$). In summary, team decisions based on the second responses weighted by the BCI confidence provided the highest and statistically better group accuracies over all methods and team sizes.

Confidence estimates provided by the BCI were more accurate in predicting the accuracy than the confidence reported by the participants alone (figure 3(B)). When averaging confidence, accuracy, and BCI confidence within participants, the reported confidence did not correlate with their accuracies in either the first ($R_S = 0.07$, Cohen's $d = 1.27$, $p = 0.80$) or second ($R_S = -0.26$, Cohen's $\underline{d} = 1.45$, $p = 0.37$) decisions. Instead, the average BCI confidence significantly correlated with the individual average accuracy in both the first ($R_S = 0.87$, Cohen's $d = -0.31$, $p = 4.7 \times 10^{-5}$) and second ($R_S = 0.67$, Cohen's $d = 0.009$, $p = 0.01$) decisions.

Figure 2. Behavioral outcomes. (A) Accuracy of human participants in the first (pre-feedback) and second (post-feedback) decisions compared to the artificial agent. (B) Task learning during the first and second decisions of humans and agent over different blocks of the experiment. (C) Response time vs. correctness of human participants in the first and second decisions. (D) Response times and correctness of human participants during the first and second decisions. (E) Subjectively reported trust of human participants in the artificial agent at the end of each block. Error bars show standard error of the mean across participants. *** $= p \leqslant 0.001$; * $= p \leqslant 0.05$; ns $=$ non-significant.

3.2. Neural markers of objective and subjective confidence

Participants' correct decisions were associated with activity of the bilateral superior parietal lobule (SPL) and the left visual cortex (figure 4(A-I), table 1), whereas the temporal neural markers were found predominantly in the left hemisphere (figure 4(B-I)), peaking at 710 ms after the stimulus onset (figure 4(B-III)). Correct trials were characterized by lower amplitude of EEG signal compared to incorrect trials (all at $p \leqslant 0.05$), with spectral information found at low frequencies (1–10 Hz) approximately 1100 ms after the stimulus onset in frontal-temporal electrodes F3–T7 and parietal-occipital electrodes Pz to O1, Oz and O2 (figure 4(C-I)). Neural oscillations were significantly different between correct and incorrect trials in the frontal-temporal beta band (W $=$ 12.0, Cohen's d $=$ 0.18, $p =$ 0.009) and the parietal-occipital delta (W $=$ 2.0, Cohen's d $=$ -0.32, $p =$ 0.0004) and beta (W $=$ 0.0, Cohen's d $=$ 0.49, $p =$ 0.0001) bands (figure 4(C-III)).

Participants' confidence was characterized by activity of the left middle frontal gyrus (MFG),

premotor cortex, SPL, right caudate nucleus, and cerebellum (lobule VII) (figure 4(A-II), table 1). The corresponding temporal neural markers were located predominantly in the parietal-occipital electrodes P1, Pz, P2, O1, Oz, and O2, as well as frontal-temporal electrodes F3, F7, and FT7 (figure 4(B-II)), peaking at approximately 1200 ms after the stimulus onset (figure 4(B-III)). Confident trials were characterized by higher amplitude of the EEG signal compared to non-confident trials (all at $p \leqslant 0.05$) (figure 4(B-IV)). Neural oscillations were significantly different between confident and non-confident trials in the frontal-temporal alpha band (W $=$ 10.0, Cohen's d $=$ 0.31, $p =$ 0.005) (figure 4 C-IV).

3.3. Neural markers of trust

Trust estimated from behavioral responses showed a significant increase over time (figure 5(B), Friedman's test W $=$ 0.42, Q $=$ 29.2, $p =$ 2.1×10^{-5}) and correlated with trust reported by participants at the end of each block ($R_S =$ 0.45, Cohen's d $=$ 1.06, $p =$ 1.6×10^{-5}).

Participants distrusting the agent showed increased activity in the left SPL compared to

Figure 3. Decoders of confidence. (A) Team performance. Average accuracy of groups of different sizes when making decisions using standard majority, weighted majority using reported confidence, and weighted majority using confidence decoded from EEG signals by the BCI, and either the first or second decision provided by the participants. (B) BCI confidence significantly correlated with accuracy in both first and second decisions, as opposed to reported confidence that did not. Dots represents the reported (pink) and BCI (blue) confidence and accuracy in the first (left) or second (right) decision of each participant. Solid lines represent linear regression models. Dashed lines represent ideal correlation.

participants who trusted the agent (figure 5(C), table 1). Similar regional temporal changes were found in the EEG signal, peaking at 880 ms after the presentation of feedback. Temporal neural markers of trust were found in the central (C1, Cz, CPz, FCz, C2) and parietal-occipital electrodes (Pz, O1,

Oz, O2), peaking between 900 and 1050 ms after feedback presentation (figure 5(D)). The distrust trials were characterized by higher EEG signal amplitude compared to the trust trials (all $p \leqslant 0.05$) (figure 5(D)). Neural oscillations were significantly different between trust and distrust trials in the

Figure 4. Neural markers of accuracy and subjective confidence. (A) Statistically significant differences in brain activity between (I) correct and incorrect trials and (II) confident and non-confident trials (II), as identified from fMRI data analysis. Color bars represent the *t* score at corrected $p \leqslant 0.05$. (B) Scalp maps represent median EEG activity and corresponding Wilcoxon *p*-values for (I) correct and incorrect trials and (II) confident and non-confident trials at different time points. Median EEG activity recorded at characteristic electrode locations across trials of different degree of (III) accuracy and (IV) subjective confidence. Asterisks indicate significantly different time points at corrected $p \leqslant 0.05$. (C) Wilcoxon *p*-values of time-frequency analysis compare EEG epochs of different degrees for (I) accuracy and (II) subjective confidence. Median power in delta (1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), and beta (12–30 Hz) EEG bands between 700 and 1400 ms for accuracy (III) and between 900 and 1400 ms for subjective confidence (IV). * denotes statistical significance at corrected $p \leqslant 0.0125$. ns = not significant; PreM = premotor cortex; Cbl-VII = cerebellar lobule VII; SPL = superior parietal lobule; V1 = primary visual cortex; CN = caudate nucleus; MFG = middle frontal gyrus.

Table 1. Neural markers of accuracy, confidence, and trust used by BCI for augmenting decision-making.

Structural area	Cluster peak z value	Cluster peak x, y, z	Cluster size mm^3
Accuracy (Correct > Incorrect)			
L superior parietal lobule (7 A), extending to L inferior parietal cortex (PGa)	3.6	−30, −64, 43	479
L primary visual cortex (V1), extending to L secondary visual cortex (V2)	3.6	−21, −51, −4	236
R superior parietal lobule (7PC), extending to R intraparietal suculus	3.2	35, −51, 47	229
Subjective confidence (Confident > Not confident)			
L middle frontal gyrus	3.9	−33, 44, −2	2622
L premotor cortex, extending to L supplementary motor area and primary motor cortex	4.4	−30, −33, 49	2011
L superior parietal lobule (7 A and 7PC)	4.1	−18, −66, 47	1977
R caudate nucleus	3.9	6, 8, 10	1964
R cerebellum (lobule VI)	4.0	17, −63, −10	1326
R cerebellum (lobule VIIa—Crus 2 and 1)	4.7	30, −70, −34	1221
Trust (Trust < Distrust)			
L superior parietal lobule (7 A and 7PC), extending to L interior parietal cortex (PGa)	3.8	−19, −76, 43	432
L superior parietal lobule (7 A)	3.5	−27, −66, 58	202
Abbreviations: L—Left; R—Right			

Figure 5. Neural markers of trust. (A) EEG epochs were split into 'Trust' and 'Distrust' trials based on the participant's first and second decisions and the agent's decision. The percentage of trials in each category is indicated in the decision tree. (B) The behavioral model of trust (teal) and subjective ratings of trust (black) across participants. (C) Statistically significant differences in brain activity between trust and distrust trials as identified from fMRI and EEG recordings. For fMRI, the color bar represents the t score at $p \leqslant 0.05$; for EEG, the color bar represents the Wilcoxon p-value. (D) Median EEG activity recorded at a characteristic electrode location across trust and distrust trials, starting at the presentation of the feedback screen. * indicates the time points at which trust and distrust distributions were significantly different at Wilcoxon $p \leqslant 0.05$. Epochs were subsampled to 250 Hz and balanced across conditions for visualization purposes. (E) Median power in central theta (4–8 Hz) band between 100 and 450 ms after feedback for trust and distrust trials. *denotes statistical significance at corrected $p \leqslant 0.0125$.

central theta band (W $=$ 12.0, Cohen's $d = 0.21$, $p = 0.009$) (figure 5(E)).

4. Discussion

Using a combined paradigm of behavioral evaluations and simultaneous EEG/fMRI recordings during a realistic pandemic scenario, we developed a collaborative BCI that accurately estimated the probability of a correct decision from the neural signal, significantly augmenting team decision-making accuracy in identification of danger zones.

Our behavioral data demonstrate that receiving feedback about a decision from the artificial agent increases the performance of the entire team but not its individual members. The most accurate teams were those whose members were equipped with the BCI to estimate their confidence in each decision from neural activity and adjust their final decision following artificial agent's feedback. Among these, the improvements brought by the BCI over the standard majority depended on the size of the team and were greater for larger teams. This finding is consistent with previous studies showing that decision-making strategies of larger teams are generally more accurate than those of smaller ones (Surowiecki 2004, Valeriani and Poli 2019). While team-based decision-making performance is monotonically increasing with the team size, it is particularly interesting that the BCI provides the biggest improvement to the smaller teams. These data also suggest that, for smaller teams, a better decoding of confidence through the BCI could compensate the absence of artificial agent's feedback.

Despite these improvements in team decision-making, participants by and large underestimated their trust in the artificial agent. However, modeling trust as a compliance and agreeableness problem revealed that participants' trust in the artificial agent significantly increased over time. These opposing findings from the subjective vs. modeling estimates of trust indicate that these measures may underpin different cognitive constructs. Alternatively, humans may require more time to process information to correctly evaluate and rely on their level of trust toward the artificial agent. Importantly, these findings suggest that trust development as perceived by participants did not interfere with the ability of the BCI to decode confidence.

To augment team decision-making during the critical pandemic scenario, the BCI was built to use the neural markers of accuracy, subjective confidence, and trust. Specifically, we used fMRI data to validate that EEG features used by the BCI capture the neural correlates of accuracy. Spatially, the accuracy markers involved SPL and visual cortex, while their temporal signature emerged in the parietal-occipital areas as early as 200 ms after the stimulus onset. The accuracy neural markers relied on significantly

modulated parietal-occipital beta and delta rhythms, as well as the frontal-temporal beta rhythm. These findings point to the importance of early engagement of visual processing and its integration with parietal and prefrontal activity for decision forming. Intersecting with accuracy, the spatial markers of subjective confidence also engaged the SPL but extended further into premotor, middle frontal, basal ganglia, and cerebellar regions, tapping into the planning, reward, and sensorimotor integration processes. In the frequency domain, subjective confidence modulated the parietal-occipital alpha rhythm. In line with our findings and relevant to the employed decision-making task, alpha and beta oscillatory activity has been previously related to various aspects of perceptual judgement and working memory, with alpha modulations setting the state of the system and beta activity encoding the stimulus properties in a context-dependent manner, both leading to the final decision outcome (Haegens *et al* 2011, 2017, Samaha *et al* 2017, Spitzer and Haegens 2017, Fischer *et al* 2018). Thus, accuracy and subjective confidence in the context of collaborative BCI engaged parallel, spatially distributed and temporally distinct neural circuits, with the former being focused on incorporating perceptual information processing and the latter involving action planning and executive operations during decision-making.

With respect to neural representations of trust, prior studies have implicated a much wider network of brain regions, including the prefrontal, paracingulate, and insular cortices, as well as the septal and ventral tegmental areas (Krueger *et al* 2007, van den Bos *et al* 2009, Fleming *et al* 2018). Notably, these studies were conducted during highly controlled games designed to specifically promote participants' trust to win. Conversely, our experimental setup used the collaborative BCI and incorporated trust in the artificial agent during a decision-making scenario. We found that, similar to accuracy and subjective confidence, the neural marker of trust was also confined to the SPL, with distrust in the artificial agent eliciting greater activity in this region. The SPL has been implicated in visual search efficiency, topographical representation of visual scenes, and spatial orientation (Corbetta *et al* 1995, Gogos *et al* 2010, Chen *et al* 2013, Lester and Dassonville 2014, Bueicheků *et al* 2015) and shown to be involved in switching attention between competing percepts and manipulating information within the working memory (Koenigs *et al* 2009, Kanai *et al* 2010, Megumi *et al* 2015), all being important contributors to perceptual decision-making. Participants' trust in the artificial agent was further characterized by the modulation of the power in the parietal-occipital theta band, suggesting specialized temporal involvement of this region in higher levels of conflict (trust/distrust) processing and resolution. Lower power in the theta band has been associated with increased feedback processing

in decision-making (Cohen *et al* 2009) and increased attention (van Driel *et al* 2012), which may be implicated in our experimental setting as a need for taking into account the decision of the artificial agent when participants trusted it. Collectively, our data suggest that the SPL emerges as a pivotal region of shared control of confidence and trust by flexibly modulating its activity and engaging premotor, prefrontal, visual and subcortical areas for the participation in various components of BCI-augmented team decision-making.

Significant improvements in team performance provided by BCI confidence over subjective confidence suggest that BCIs are more accurate in estimating confidence than humans alone. Future studies should explore the effects of having humans collaborate with another human via a machine-mediated confidence system. For example, an artificial agent could be using decisions and BCI confidence estimates of a human who performed the task previously. This paradigm may also help disentangle the biases in trust development between the human-machine and human-human teams.

In summary, we developed a multimodal BCI that assisted human-artificial agent teams in improving their decision-making in a pandemic scenario. The unique neural pattern of accuracy markers combined with spatial-temporal transformations enabled our BCI to accurately decode confidence from EEG recordings and deliver significant improvements of team performance. The BCI-decoded confidence better correlated with the decision accuracy than the confidence subjectively reported by the participants. Our findings demonstrate that these BCIs may be used in the operational settings characterized by diverse human-artificial agent collaborations for augmented team performance leading to optimal decisions in critical situations.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

Conflict of interest

None relevant to this study for all authors.

Contributions

Designed research—D V, K S; Performed research—D V, L C O, A W, A H S; Analyzed data—D V, L C O; Wrote the paper—D V, L C O, K S

Code availability

Codes are available on https://github.com/simonyanlab/DoD-BARI-BCI-Data-Repository. The Material and Methods and Figures contain all details on used codes needed to reproduce the results.

Significance statement

This research developed the first multimodal, collaborative brain-computer interface (BCI) to disentangle the spatial-temporal neural markers of decision confidence and trust between humans and an artificial agent and to train confidence decoders for augmenting the collaborative team accuracy during complex decisions. The results of this study lay a foundation for the future development of consumer-grade BCIs that can be used in real-life situations.

ORCID iDs

Davide Valeriani ⬤ https://orcid.org/0000-0001-9866-0063
Kristina Simonyan ⬤ https://orcid.org/0000-0001-7444-0437

References

Allen P J, Josephs O and Turner R 2000 A method for removing imaging artifact from continuous EEG recorded during functional MRI *Neuroimage* **12** 230–9

Bang D and Fleming S M 2018 Distinct encoding of decision confidence in human medial prefrontal cortex *Proc. Natl Acad. Sci. USA* **115** 6082–7

Boldt A and Yeung N 2015 Shared neural markers of decision confidence and error detection *J. Neurosci.* **35** 3478–84

Bueichekú E, Ventura-Campos N, Palomar-García M, Miró-Padilla A, Parcet M A and Ávila C 2015 Functional connectivity between superior parietal lobule and primary visual cortex "at rest" predicts visual search efficiency *Nat. Neurosci.* **5** 517–26

Chen J, Feng T, Shi J, Liu L and Li H 2013 Neural representation of decision confidence *Behav. Brain Res.* **245** 50–57

Cohen M X, Elger C E and Fell J 2009 Oscillatory activity and phase-amplitude coupling in the human medial frontal cortex during decision making *J. Cogn. Neurosci.* **21** 390–402

Corbetta M, Shulman G L, Miezin F M and Petersen S E 1995 Superior parietal cortex activation during spatial attention shifts and visual feature conjunction *Science* **270** 802–5

Drnec K, Marathe A R, Lukos J R and Metcalfe J S 2016 From trust in automation to decision neuroscience: applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction *Front. Hum. Neurosci.* **10** 290

Evans A M and Revelle W 2008 Survey and behavioral measurements of interpersonal trust *J. Res. Pers.* **42** 1585–93

Fernandez-Vargas J, Tremmel C, Valeriani D, Bhattacharyya S, Cinel C, Citi L and Poli R 2021 Subject- and

task-independent neural correlates and prediction of decision confidence in perceptual decision making *J. Neural Eng.* **18** PMID: 33780913

Ferree T C, Luu P, Russell G S and Tucker D M 2001 Scalp electrode impedance, infection risk, and EEG data quality *Clin. Neurophysiol.* **112** 536–44

Fischer A G, Nigbur R, Klein T A, Danielmeier C and Ullsperger M 2018 Cortical beta power reflects decision dynamics and uncovers multiple facets of post-error adaptation *Nat. Commun.* **9** 5038

Fleming S M and Lau H C 2014 How to measure metacognition *Front. Hum. Neurosci.* **8** 443

Fleming S M, van der Putten E J and Daw N D 2018 Neural mediators of changes of mind about perceptual decisions *Nat. Neurosci.* **21** 617–24

Gherman S and Philiastides M G 2018 Human VMPFC encodes early signatures of confidence in perceptual decisions *Elife* **7** e38293

Gogos A, Gavrilescu M, Davison S, Searle K, Adams J, Rossell S L, Bell R, Davis S R and Egan G F 2010 Greater superior than inferior parietal lobule activation with increasing rotation angle during mental rotation: an fMRI study *Neuropsychologia* **48** 529–35

Gramfort A *et al* 2013 MEG and EEG data analysis with MNE-Python *Front. Neurosci.* **7** 267

Haegens S, Nácher V, Luna R, Romo R and Jensen O 2011 α-Oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking *Proc. Natl Acad. Sci. USA* **108** 19377–82

Haegens S, Vergara J, Rossi-Pool R, Lemus L and Romo R 2017 Beta oscillations reflect supramodal information during perceptual judgment *Proc. Natl Acad. Sci. USA* **114** 13810–5

Kanai R, Bahrami B and Rees G 2010 Human parietal cortex structure predicts individual differences in perceptual rivalry *Curr. Biol.* **20** 1626–30

Koenigs M, Barbey A K, Postle B R and Grafman J 2009 Superior parietal cortex is critical for the manipulation of information in working memory *J. Neurosci.* **29** 14980–6

Krueger F, McCabe K, Moll J, Kriegeskorte N, Zahn R, Strenziok M, Heinecke A and Grafman J 2007 Neural correlates of trust *Proc. Natl Acad. Sci. USA* **104** 20084–9

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 Comput. Interfaces* **6** 162–77

Lester B D and Dassonville P 2014 The role of the right superior parietal lobule in processing visual context for the establishment of the egocentric reference frame *J. Cogn. Neurosci.* **26** 2201–9

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

Megumi F, Bahrami B, Kanai R and Rees G 2015 Brain activity dynamics in human parietal regions during spontaneous switches in bistable perception *Neuroimage* **107** 190–7

Meyen S, Sigg D M B, Luxburg U V and Franz V H 2021 Group decisions based on confidence weighted majority voting *Cogn. Res. Princ. Implic.* **6** 18

Moore D and Cain D 2007 Overconfidence and underconfidence: when and why people underestimate (and overestimate) the competition *Organ. Behav. Hum. Decis. Process.* **103** 197–213

Navajas J, Hindocha C, Foda H, Keramati M, Latham P E and Bahrami B 2017 The idiosyncratic nature of confidence *Nat. Hum. Behav.* **1** 810–8

Niazy R K, Beckmann C F, Iannetti G D, Brady J M and Smith S M 2005 Removal of FMRI environment artifacts from EEG data using optimal basis sets *Neuroimage* **28** 720–37

Park J-G and Lee J 2014 Knowledge sharing in information systems development projects: explicating the role of dependence and trust *Int. J. Proj. Manage.* **32** 153–65

Pereira M, Faivre N, Iturrate I, Wirthlin M, Serafini L, Martin S, Desvachez A, Blanke O, Van De Ville D and Millan J D R 2020 Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging *Proc. Natl Acad. Sci. USA* **117** 8382–90

Pescetelli N and Yeung N 2021 The role of decision confidence in advice-taking and trust formation *J. Exp. Psychol. Gen.* **150** 507–26

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

Platt J 1999 Probabilistic outputs for SVMs and comparisons to regularized likehood methods *Advances in Large Margin Classifiers.* (Cambridge, MA: MIT Press)

Pouget A, Drugowitsch J and Kepecs A 2016 Confidence and certainty: distinct probabilistic quantities for different goals *Nat. Neurosci.* **19** 366–74

Rivet B, Souloumiac A, Attina V and Gibert G 2009 xDAWN algorithm to enhance evoked potentials: application to brain–computer interface *IEEE Trans. Biomed. Eng.* **56** 2035–43 PMID: 19174332

Samaha J, Iemi L and Postle B R 2017 Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy *Conscious. Cogn.* **54** 47–55

Shekhar M and Rahnev D 2018 Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition *J. Neurosci.* **38** 5078–87

Spitzer B and Haegens S 2017 Beyond the status quo: a role for beta oscillations in endogenous content (re)activation *eNeuro* **4** ENEURO.0170–17.2017

Surowiecki J 2004 *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* (New York: Doubleday)

Valeriani D, Cinel C and Poli R 2017 Group augmentation in realistic visual-search decisions via a hybrid brain-computer interface *Sci. Rep.* **7** 7772

Valeriani D and Poli R 2019 Cyborg groups enhance face recognition in crowded environments *PLoS One* **14** e0212935

van den Bos W, van Dijk E, Westenberg M, Rombouts S A and Crone E A 2009 What motivates repayment? Neural correlates of reciprocity in the Trust Game *Soc. Cogn. Affect. Neurosci.* **4** 294–304

van Driel J, Ridderinkhof K R and Cohen M X 2012 Not all errors are alike: theta and alpha EEG dynamics relate to differences in error-processing dynamics *J. Neurosci.* **32** 16795–806