frontiers in Virtual Reality

TYPE Original Research
PUBLISHED 13 November 2026
DOI 10.3389/frvir.2026.1971503

Check for updates

OPEN ACCESS

EDITED BY
Beatrice Aruanno,
University of Pisa, Italy

REVIEWED BY
Jesse A. Mark,
The University of Western Ontario, Canada
Gaang Lee,
University of Alberta, Canada

*CORRESPONDENCE
Christopher Baker,
c.baker@qub.ac.uk

RECEIVED 18 August 2026
REVISED 18 September 2026
ACCEPTED 30 October 2026
PUBLISHED 13 November 2026

CITATION
Baker C, Hinton S, Nijjar A, Reed T and Fairclough S (2026) Decoding decision correctness from EEG under high cognitive workload in virtual reality: implications for collaborative brain-computer interface teams.
Front. Virtual Real. 7:1971503.
doi: 10.3389/frvir.2026.1971503

COPYRIGHT
© 2026 Baker, Hinton, Nijjar, Reed and Fairclough. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Decoding decision correctness from EEG under high cognitive workload in virtual reality: implications for collaborative brain-computer interface teams

Christopher Baker[1]*, Stephen Hinton[2], Akashdeep Nijjar[3], Thomas Reed[4] and Stephen Fairclough[2]

[1]School of Electronics, Electrical Engineering and Computer Science Queen's University Belfast, Belfast, United Kingdom, [2]School of Psychology, Liverpool John Moores University, Liverpool, United Kingdom, [3]School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom, [4]Defence Science and Technology Laboratory, Salisbury, United Kingdom

Introduction: Collaborative Brain-Computer Interfaces (cBCIs) offer a promising mechanism to augment team decision-making, but existing approaches rely exclusively on evidence available only after a decision has been made and reported, such as reaction time or stated confidence. This limits their use to explaining or discounting a decision after the fact, rather than informing a teams response before it is finalised. We tested whether spatial-covariance EEG features could instead provide a genuinely pre-emptive signal of an operators decision correctness, available within the response window itself, and whether such a signal depends on cognitive workload.

Methods: Using a continuous virtual reality target-detection task, participants (N = 23) completed a within-subject visual workload manipulation (High vs. Low), implemented via a Degraded Visual Environment.

Results: A Riemannian tangent-space classifier, trained on a strictly motor-safe EEG window preceding each response, decoded Correct from Incorrect decisions with above-chance accuracy under High Workload (61.4%, p = .0014, N = 23) but not reliably under Low Workload (N = 17), where near-ceiling behavioural accuracy left too few error trials for valid classification. Ablation analysis showed the signal was carried predominantly by posterior occipito-parietal channels, consistent with a perceptual rather than motor origin. At the team level, weighting votes by this pre-emptive neural signal, available before a response is committed, produced substantial accuracy gains on contested (evenly-split) trials under High Workload (57% to 88% as team size increased from 2 to 16), but was actively detrimental under Low Workload.

Discussion: Critically, this advantage held even against post-hoc behavioural signals: confidence was the strongest single team-level signal overall, but by definition cannot inform a decision still in progress, whereas the neural signal can. These findings indicate that EEG-based decision-reliability signals are not a general-purpose team augmentation tool, but a workload-conditional one, with clear implications for when and how cBCI systems should be deployed in operational teams.

KEYWORDS
brain-computer interface, cognitive workload, EEG decoding, pre-emptive signal, Riemannian geometry, team decision-making, virtual reality

Introduction

Modern operational environments increasingly rely on teams of human operators to make rapid perceptual decisions under demanding conditions, monitoring surveillance feeds, screening medical images, or detecting threats in degraded visibility (Cabrera et al., 2023; Canonico et al., 2019; Dehais et al., 2022; Goldfarb and Lindsay, 2022; Koopman and Zammit-Mangion, 2024). The statistical advantage of aggregating such teams, often termed the "wisdom of crowds", is fundamentally predicated on the independence and reliability of individual judgments (Galton, 1907; Hamada et al., 2020; Kerr and Tindale, 2004). However, cognitive workload systematically degrades individual perceptual reliability (Kosch et al., 2023; Plainis and Murray, 2002; Wickens, 2008), and conventional team aggregation methods, whether simple majority vote or approaches weighted by reaction time or past performance, implicitly assume that an operator's explicit behavioural output remains a trustworthy proxy for their underlying perceptual state (Arshad et al., 2015; Bhattacharyya et al., 2019; Cinel et al., 2022; Poli et al., 2014; Woolley et al., 2010). Under sustained high workload, this assumption may not hold.

Contemporary models of metacognition suggest that the sensory evidence contributing to a perceptual decision can be computationally dissociated from the process governing its behavioural report (Fleming and Lau, 2014; Wokke et al., 2020; 2017). Electrophysiological work demonstrates that neural markers of decision correctness, and even of error commission, can be detected independently of, and sometimes prior to, an operator's own explicit awareness or report (Di Gregorio et al., 2020; Murayama et al., 2016; Sadras et al., 2023; Schnuerch et al., 2024; Yeung and Summerfield, 2012). This raises the possibility that an operator's true decision reliability is only partially observable through the behavioural channels, such as reaction time or stated confidence, that current team aggregation methods rely on. Critically, these behavioural channels are also only available after a decision has been reported, meaning conventional aggregation can, at best, discount an unreliable vote after the fact. A signal available earlier, within the decision process itself, would instead allow team aggregation to weight a vote appropriately at the moment it is cast.

Collaborative Brain-Computer Interfaces (cBCIs) have been proposed as a mechanism to give team aggregation access to this otherwise unobservable channel, using neural signals to weight or inform collective decisions directly (Bhattacharyya et al., 2021; Salvatore et al., 2022; Valeriani et al., 2022; 2019; 2017). In continuous, self-paced, ecologically valid environments such as virtual reality, where operators visually acquire a target before any formal decision cue appears, traditional time-domain analyses are vulnerable to substantial latency jitter between stimulus onset and neural response (Hassan et al., 2024; Quitadamo et al., 2017; Zhou et al., 2022). Spatial covariance approaches using Riemannian geometry offer a robust alternative, capturing the relational structure of oscillatory brain activity across the full electrode array rather than depending on precise temporal alignment (Barachant et al., 2012; Congedo et al., 2017; Yger et al., 2018).

The present study applies this approach to a specific and operationally relevant decoding target: not what an operator saw, but whether their judgment about it was correct. This distinction matters because team aggregation ultimately depends on knowing which operators to trust on a given trial, not on independently verifying the world state each operator observed. Using a continuous virtual reality target-detection task under a within-subject manipulation of cognitive workload, we test, first, whether a Riemannian spatial-covariance classifier, trained on a strictly motor-safe EEG window, can decode decision correctness; second, whether this signal is conditional on workload, and what its likely neural origin is; and third, whether it provides real, translatable value when aggregated across simulated teams, evaluated specifically against, not merely alongside, the behavioural signals (confidence, reaction time) that conventional aggregation already has access to. In doing so, we establish the specific operational conditions under which such a signal should be deployed.

Methods

Participants

Twenty-three participants (Mean age ±SD = 24.8 ± 6.0; 10 Female) took part in the study, following exclusion of participants during preliminary quality control for persistent LabStreamingLayer synchronisation failures or excessive baseline EEG artefacts. All participants reported normal or corrected-to-normal vision, no history of neurological disorders, and no susceptibility to VR-induced motion sickness. Written informed consent was obtained from all individuals. The experimental protocol received favourable opinion from the UK Ministry of Defence Research Ethics Committee (MoDREC, Reference: RQ0000037929), and all procedures were conducted in accordance with the Declaration of Helsinki.

Study design and VR target-detection task

Participants completed a within-subject repeated-measures design within a continuous Virtual Reality (VR) target-detection task. Participants viewed a high-fidelity 3D simulation rendered in Unreal Engine 5 via a Varjo Aero Head-Mounted Display (Varjo Technologies, Helsinki, Finland), depicting the viewpoint of a drone traversing a landscape. Participants discriminated between continuously appearing 3D models, categorised as Non-Targets and Targets (Figures 1A,B). A targeting reticle locked onto each stimulus for 2500 ms, prompting a physical joystick button press before the reticle disappeared (Figure 2).

Cognitive workload was manipulated within-subject across sessions by simulating a Degraded Visual Environment: High Workload sessions reduced ambient lighting by 50% and lowered the solar angle relative to Low Workload sessions, which used standard daylight conditions (Figures 3A,B). Visual degradation manipulations of this kind are an established method for manipulating cognitive workload in human factors and neuroergonomics research (Biernacki and Lewkowicz, 2024; Dehais et al., 2022; Yu et al., 2015), validated in EEG paradigms specifically via both subjective (NASA-TLX) and physiological (P300 attenuation) corroboration (Yu et al., 2015). We discuss

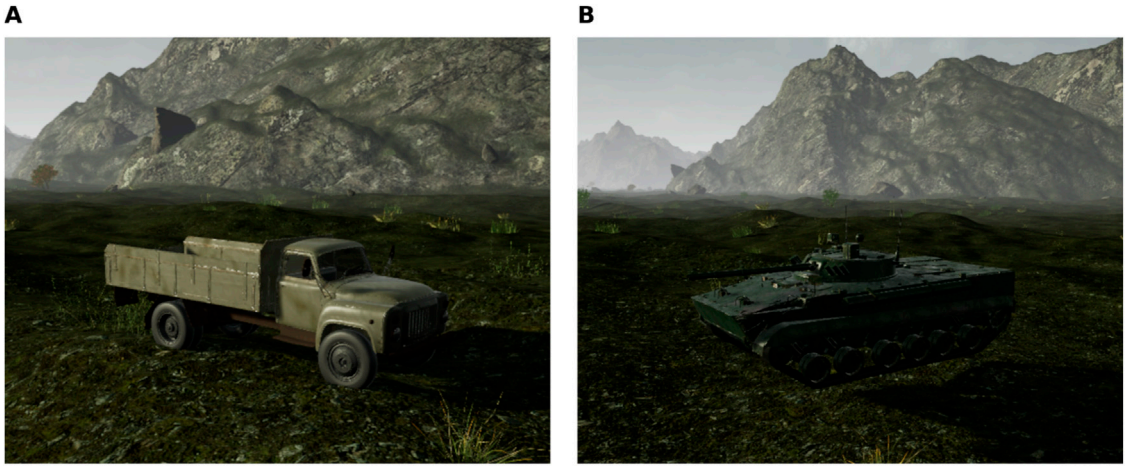

FIGURE 1
Virtual Reality target detection task stimuli. **(A)** Example of a Non-Target stimulus. **(B)** Example of a Target stimulus.

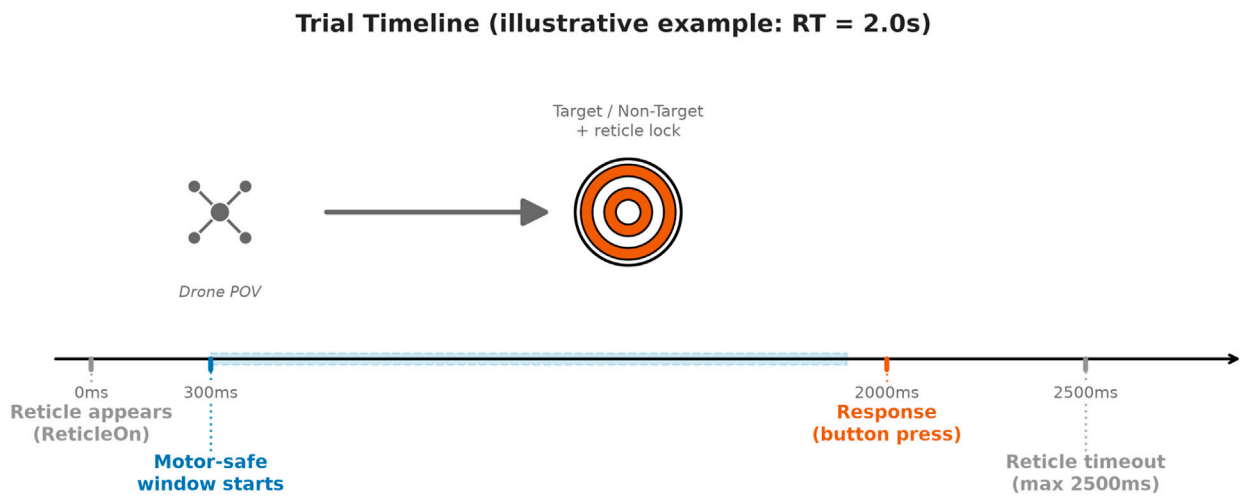

FIGURE 2
Trial timeline. A targeting reticle locks onto a Target or Non-Target stimulus at 0 ms and remains locked for a maximum of 2500 ms, within which the participant executes a joystick button press. All EEG features are extracted from a motor-safe window beginning 300 ms post-reticle and ending 100 ms before each trial's own response time.

the relationship between this manipulation and cognitive workload as a construct, and the absence of independent subjective or physiological validation, in the Limitations.

Data acquisition and EEG signal processing

Continuous EEG data were recorded at 500 Hz using a 32-channel LiveAmp system (Brain Products GmbH) (Figure 4). Behavioural data and VR events were synchronised via LabStreamingLayer (LSL) to ensure sub-millisecond precision (Dirican and Göktürk, 2011; Kothe et al., 2024). EEG data were processed offline using MNE-Python (Gramfort et al., 2013). Following band-pass (0.1 to 30 Hz) and notch (50 Hz) filtering, Independent Component Analysis (ICA) was computed and components reflecting ocular artefacts were manually rejected (Ablin et al., 2018) Data were segmented into epochs (−200 ms to +800 ms relative to reticle onset).

Motor-safe feature extraction and classification

To decode decision correctness independently of motor-execution artefacts, features were extracted from a strictly motor-safe window: from 300 ms post-reticle-onset to 100 ms before each trial's own response time, with a minimum required window duration of 150 ms; trials not meeting this criterion were excluded (Figure 5). The 300 ms start point was chosen to exclude the stimulus-locked evoked response elicited by the reticle's appearance, a response present in all trials regardless

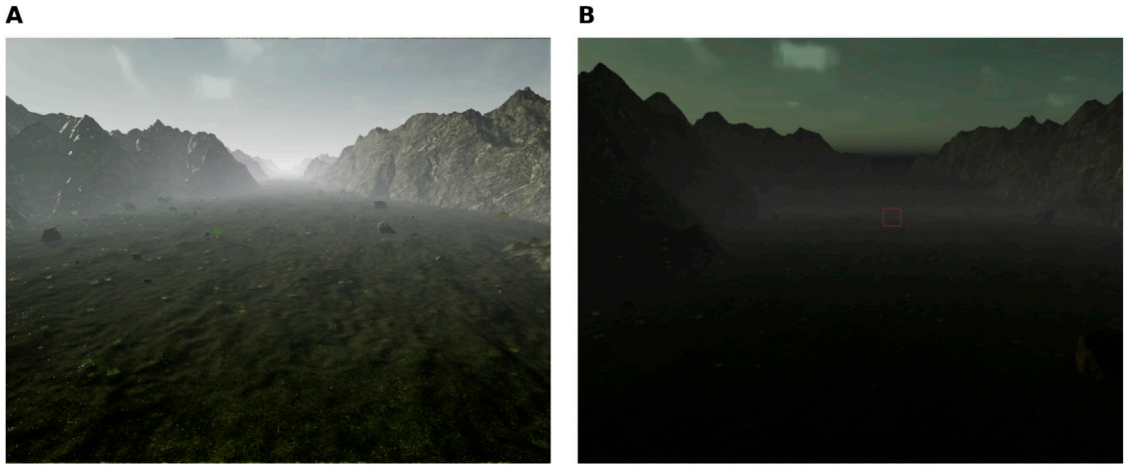

FIGURE 3
Environmental Workload Manipulations. **(A)** Low Workload standard daylight condition. **(B)** High Workload simulating a Degraded Visual Environment (DVE) via a 50% reduction in lighting.



FIGURE 4
Experimental setup demonstrating a participant wearing the active EEG cap and VR Head-Mounted Display.

of eventual decision correctness (see Results, 'The Reticle Elicits a Genuine, Motor-Safe Neural Response'), from the covariance estimate used for classification. This exclusion criterion removed 12.61% of Correct and 2.84% of Incorrect trials under High Workload, and 49.53% of Correct and 5.43% of Incorrect trials under Low Workload; we return to the implications of this asymmetry in the Results and Discussion. Within this window, a spatial covariance matrix was estimated per trial using Ledoit-Wolf shrinkage (Ledoit and Wolf, 2004) and projected into a Euclidean tangent space via the Riemannian mean (Barachant et al., 2012; Congedo et al., 2017; Yger et al., 2018), using PyRiemann. Features were standardised and classified using logistic regression with balanced class weighting (C = 0.1) (Pedregosa et al., 2011). Model validation used leave-one-subject-out (LOSO) cross-validation, separately for High and Low Workload.

Team simulation and aggregation

Team-level performance was evaluated via bootstrap resampling (500 draws per team size and session) (Harris et al., 2020; McKinney, 2010) across team sizes of 2, 4, 8, and 16. For each simulated team, votes from individual members' recorded response directions were aggregated by unweighted majority vote, and separately by evidence-weighted vote using: (a) the pre-emptive neural signal (each member's out-of-fold classifier probability, transformed to a signed evidence value), (b) post-hoc stated confidence, and (c) post-hoc response time. To isolate trials where aggregation weighting could plausibly influence the outcome, a secondary analysis restricted to true-tie trials, defined as trials on which a simulated team's raw, unweighted vote was split within one vote of even. Concretely, for a team of size N (N = 2, 4, 8, 16), a trial was classified as a true tie if the unweighted vote split was either exactly even (N/2 versus N/2) or within a single vote of even ((N/2 − 1) versus (N/2 + 1)); both configurations were included in the true-tie analysis.

Statistical analysis

Paired t-tests assessed the workload manipulation's behavioural effect (accuracy, reaction time) and the Target-detection/Non-Target-rejection asymmetry. One-sample t-tests against chance (50%) assessed classifier accuracy across LOSO folds. Cluster-based permutation testing (5000 permutations) assessed reticle-locked EEG responses against a pre-stimulus baseline(Maris and Oostenveld, 2007). Channel-level contribution to classifier accuracy was assessed via ablation: each channel's covariance row and column were zeroed in turn, and the resulting drop in LOSO accuracy relative to the full-channel baseline was taken as that channel's importance. Signal detection theory criterion (c) and sensitivity (d′) were computed per subject and workload condition from hit and false alarm rates, using a log-linear correction for extreme rates (Hautus, 1995). Classifier discrimination was additionally assessed using balanced accuracy and ROC-AUC, computed from out-of-



FIGURE 5
The motor-safe window scales with each trial's own reaction time. Trials whose window would fall below a 150 ms minimum duration are excluded from feature extraction (top row). Illustrative reaction times shown correspond approximately to the session means for Low Workload (≈0.85 s) and High Workload (≈1.4 s), and to the reticle's 2500 ms hard timeout.

TABLE 1 **Workload manipulation check. Mean accuracy and response time under high and low workload, with paired t-tests comparing the two conditions.**

Condition			Mean accuracy			Mean RT		
High workload			0.78			1.40		
Low workload			0.94			0.81		
Metric	**T**	**Dof**	**Alternative**	**p_val**	**CI95**	**cohen_d**	**Power**	**BF10**
Accuracy (high vs. low)	−8.8373	21	Two-sided	< .0001	[-0.2–0.12]	1.5881	1	8.16E+05
RT (high vs. low)	13.3937	21	Two-sided	< .0001	[0.5 0.68]	1.8959	1	8.78E+08

TABLE 2 **Target detection versus non-target rejection accuracy, paired t-test across participants.**

Trial type					Mean accuracy		
Target detection (hit rate)					0.81		
Non-target rejection (correct rejection rate)					0.90		
T	**Dof**	**Alternative**	**p_val**	**CI95**	**cohen_d**	**Power**	**BF10**
−2.6961	22	Two-sided	0.0132	[-0.16–0.02]	0.625	0.8172	3.894

TABLE 3 Leave-one-subject-out classifier performance decoding Correct versus Incorrect decisions, by workload condition. TPR: true positive rate (correct trials correctly classified). TNR: true negative rate (incorrect trials correctly classified).

Condition	Mean_accuracy	Mean_TPR_correct	Mean_TNR_incorrect	N_Subjects	t_vs._chance	p_vs._chance
High workload	0.61	0.65	0.46	23.00	3.65	0.0014
Low workload	0.79	0.86	0.12	17.00	8.66	0.0000002

fold classifier probabilities per subject; the relationship between motor-safe window duration and classifier reliability was assessed via Pearson correlation.

Results

Workload manipulation and behavioural asymmetry

The workload manipulation produced a robust behavioural effect. Accuracy was significantly lower under High Workload than Low Workload (77.55% vs. 93.84%), and response time was significantly slower (1.40s vs. 0.81s; both $p < .0001$; Table 1), confirming the manipulation successfully degraded perceptual performance. Independently of workload, Target detection was significantly harder than Non-Target rejection ($p = .007$; Table 2), an asymmetry that recurs later in the analysis of trial difficulty.

To further characterise the nature of this behavioural effect, we computed signal detection theory criterion (c) and sensitivity (d') per subject and workload condition, using hit and false alarm rates with a log-linear correction for extreme rates (Hautus, 1995). Response criterion did not differ meaningfully between conditions (High Workload: $M = 0.161$; Low Workload: $M = 0.135$; $t(21) = 0.60$, $p = .554$, $BF10 = 0.263$, moderate evidence for no difference), indicating participants did not become more liberal or conservative under degraded visibility. By contrast, sensitivity was substantially and reliably lower under High Workload ($d' = 1.98$) than Low Workload ($d' = 3.58$; $t(21) = -12.04$, $p < .0001$, $BF10 = 1.36 \times 10^8$). This pattern, a marked reduction in discriminability with no accompanying shift in response bias, indicates the workload manipulation acted primarily on perceptual sensitivity rather than decision strategy.

The reticle elicits a genuine, motor-safe neural response

Before testing whether EEG distinguishes correct from incorrect decisions, we confirmed that the pipeline detects a genuine evoked response to the reticle at all, independent of decision correctness (Figures 3, 5). Using a reaction-time-masked analysis, in which each timepoint's average included only trials whose response had not yet occurred, cluster-based permutation testing identified significant reticle-locked activity in both High Workload (two clusters, $p = .013$ and $p = .027$, spanning approximately 0.31 to 0.80 s post-reticle) and Low Workload (one cluster, $p = .028$, spanning approximately 0.28 to 0.80 s), confirming the pipeline reliably detects real signal within the motor-safe window used for all subsequent analyses.

EEG decodes decision correctness under high, but not low, workload

A Riemannian tangent-space classifier, trained on motor-safe spatial covariance features and validated via leave-one-subject-out cross-validation, decoded Correct from Incorrect decisions with above-chance accuracy under High Workload (61.4%, TPR 64.8%, TNR 46.3%, $t(22) = 3.65$, $p = .0014$, $N = 23$; Table 3). Under Low

10.3389/frvir.2026.1971503

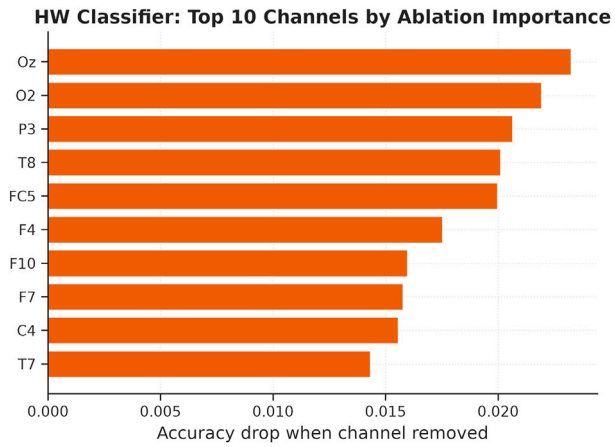

FIGURE 6
Top ten channels ranked by ablation importance (drop in leave-one-subject-out accuracy when that channel is removed), High Workload classifier.

Workload, raw classifier accuracy was numerically higher (79.4%) but not a comparable finding: true negative rate collapsed to 12.3%, indicating the classifier almost never correctly identified an incorrect trial (Table 3). This is attributable to the behavioural ceiling effect established above (93.84% Low Workload accuracy), which left too few genuine error trials for the classifier to learn a reliable decision boundary. The reduction from N = 23 to N = 17 for this analysis reflects our leave-one-subject-out validation excluding any subject whose held-out test set did not contain both Correct and Incorrect trials; combined with the motor-safe window's disproportionate exclusion of Correct trials under Low Workload (Methods), a subset of subjects retained too few genuine Incorrect trials to be included. All subsequent analyses of classifier structure and team-level value therefore focus on High Workload; the Low Workload classifier is reported but not mechanistically interpreted. Because behavioural accuracy is imbalanced between Correct and Incorrect trials, particularly under Low Workload, we additionally computed balanced accuracy and ROC-AUC as threshold-independent measures of discrimination. Balanced accuracy, the mean of the reported true positive and true negative rates, was 55.55% for High Workload and 45.85% for Low Workload. Mean ROC-AUC, computed from out-of-fold classifier probabilities per subject, was 0.553 for High Workload, significantly above chance ($t(22) = 3.12$, $p = .005$, $BF10 = 8.81$), and 0.447 for Low Workload, not significantly different from chance ($t(16) = -1.29$, $p = .217$, $BF10 = 0.504$, 95% CI [0.36, 0.53]). These convergent, imbalance-robust measures confirm genuine, if modest, decoding performance under High Workload, and the absence of reliable discrimination under Low Workload once class imbalance is accounted for. Because the motor-safe window's duration scales with each trial's own reaction time, we tested whether window duration was associated with classifier reliability. Across all High Workload trials (N = 2,773), window duration showed only a negligible association with out-of-fold classification correctness (Pearson $r = -0.086$, $p < .0001$, decisive Bayesian evidence given the large sample size) and with classifier confidence ($r = -0.056$, $p = .003$, weaker evidence), together explaining under 1% of variance in either measure. This indicates that variation in window duration, and the corresponding variation in the amount of data available for the per-trial Ledoit-Wolf covariance estimate, does not meaningfully affect classifier reliability.

To characterise which channels drove classification, we performed an ablation analysis: each channel's contribution to the spatial covariance matrix was removed in turn, and the resulting drop in leave-one-subject-out accuracy was taken as that channel's importance (Figure 6). The largest drops in accuracy occurred at occipital and parietal sites (Oz, O2, P3), with central and temporal channels (C4, T8) contributing more modestly further down the ranking. This posterior-weighted profile

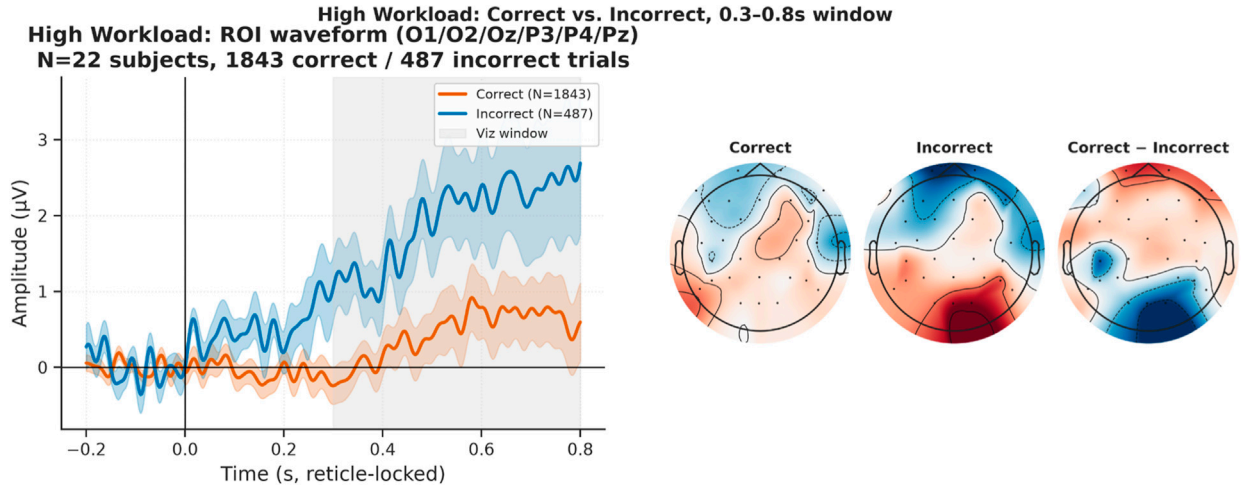

FIGURE 7
Grand-average Correct versus Incorrect waveform and scalp topography, High Workload, fixed 0.3–0.8 s visualisation window. Left: mean amplitude ±SEM at occipito-parietal sites (O1, O2, Oz, P3, P4, Pz). Right: scalp topography for Correct, Incorrect, and their difference. Low Workload is not shown, as its classifier is not mechanistically interpreted (see Results).

TABLE 4 Vote margin by Target/Non-Target status. Lower margin indicates a more contested trial (closer to an evenly split vote).

Condition	N_Trials	Mean_Margin_Target	Mean_Margin_NonTarget	t	p
High workload	147	0.48	0.69	−5.63	9.05E-08
Low workload	144	0.71	0.88	−4.78	4.29E-06

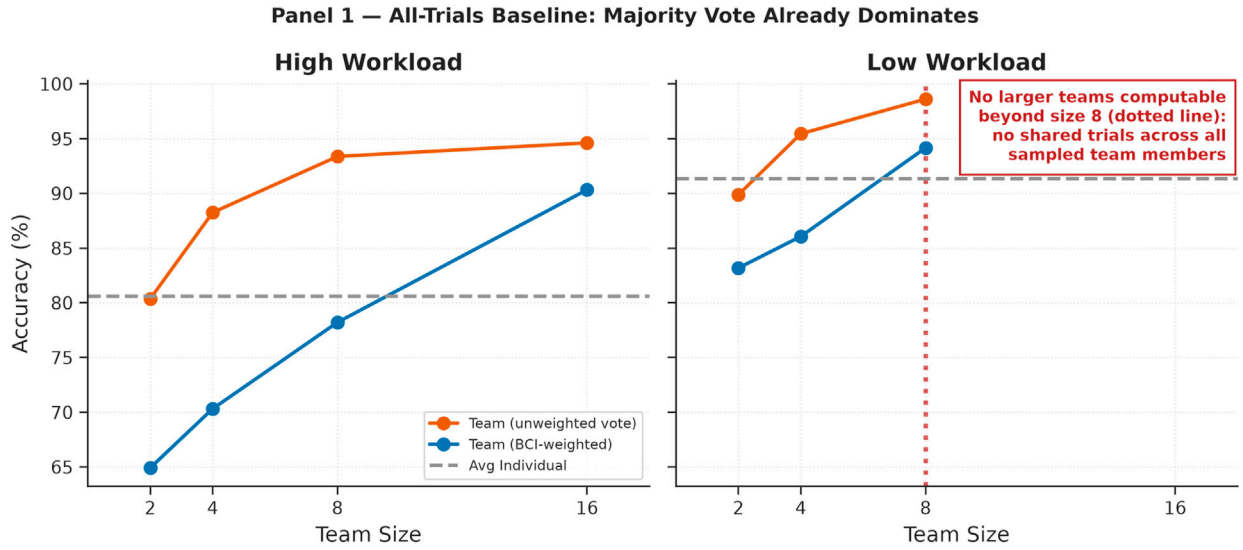

FIGURE 8
Team accuracy across all trials (no tie restriction), unweighted majority vote versus neural-weighted vote, by team size and workload. Majority vote alone already scales close to ceiling, motivating the tie-restricted analysis in Figure 9.

is consistent with a perceptual rather than motor origin for the decoded signal (Pfurtscheller and Lopes da Silva, 1999) (Figure 7), and is a more conservative and defensible characterisation than raw classifier coefficient magnitude, which we found to be diffusely distributed across channels due to L2 regularisation and did not reliably distinguish genuinely informative channels from correlated, redundant ones. We note that this ablation compares Correct against Incorrect trials within the High Workload condition itself, so it cannot be explained by the between-condition luminance difference between High and Low Workload; however, trial-level variation in visibility or perceptual demand within High Workload could still contribute to both posterior channel activity and decision correctness, a possibility our signal detection theory analysis below speaks to directly.

Trial difficulty is explained by the target/non-target asymmetry

To characterise what made a trial contested, we computed, for every trial with at least three available responses, a vote margin: the absolute value of the summed response directions divided by the number of respondents, ranging from 0 (perfectly split) to 1 (unanimous). Target trials showed significantly lower vote margins (more contested) than Non-Target trials, in both High Workload ($t = -5.63$, $p < .0001$, $N = 147$ trials) and Low Workload ($t = -4.78$, $p < .0001$, $N = 144$ trials; Table 4). This indicates that trial difficulty, measured independently via inter-rater disagreement, is substantially explained by the same Target-detection asymmetry established behaviourally, rather than reflecting an unrelated or unexplained source of difficulty.

Pre-emptive neural weighting improves team accuracy on contested trials under high workload

We next asked whether this individual-level signal, used pre-emptively to weight each team member's vote before the team's collective decision, provided value at the team level. Because majority vote alone already performs strongly on the full trial set, rising to 94%–99% accuracy by team size 16 in both workload conditions (Figure 8), any weighting scheme can only plausibly change a team's outcome on trials where the unweighted vote is tied. We therefore restricted this analysis to true-tie trials, using bootstrap-resampled simulated teams (500 draws per team size and session) at sizes of 2, 4, 8, and 16. The proportion of trials meeting this true-tie criterion decreased substantially as team size increased. Under High Workload, true ties comprised 29.3% of trials at team size 2, falling to 15.3% at team size 4, 6.6% at team size 8, and 3.4% at team size 16. Under Low Workload, true ties comprised 18.5% of trials at team size 2, 6.9% at team size 4, and 2.5% at team size 8; no shared trials were available across all sampled team members at team size 16 under Low Workload (Figure 9), so no

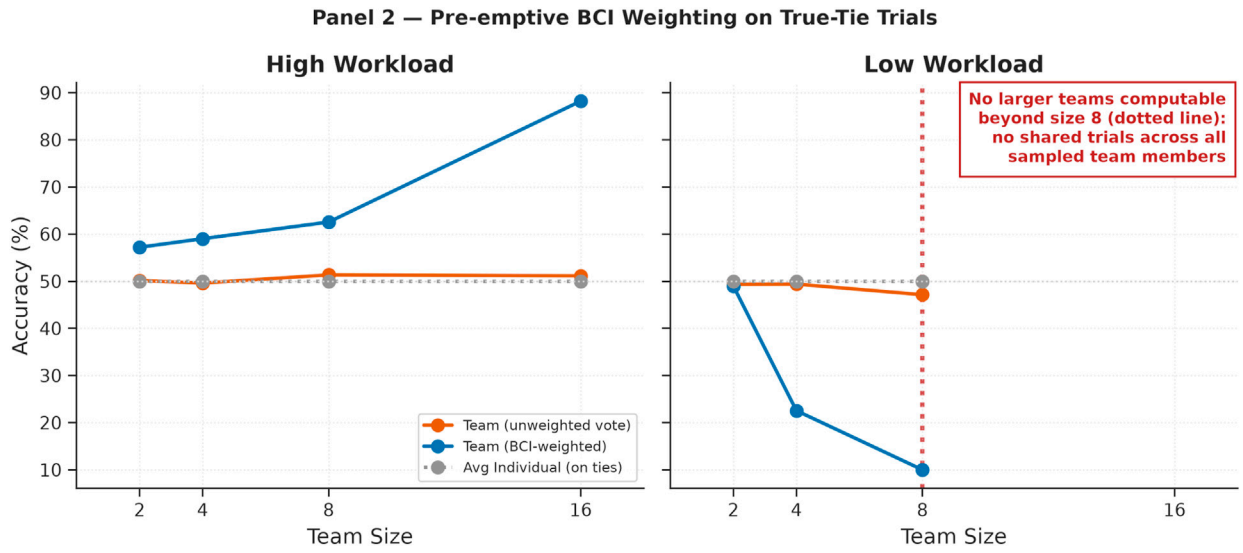

FIGURE 9
Team accuracy on true-tie trials (trials on which the unweighted vote was split within one vote of even), unweighted vote versus pre-emptive neural weighting, by team size and workload. Dotted line shows average individual accuracy conditioned on the same tie trials (≈50%, the base rate implied by a tie).

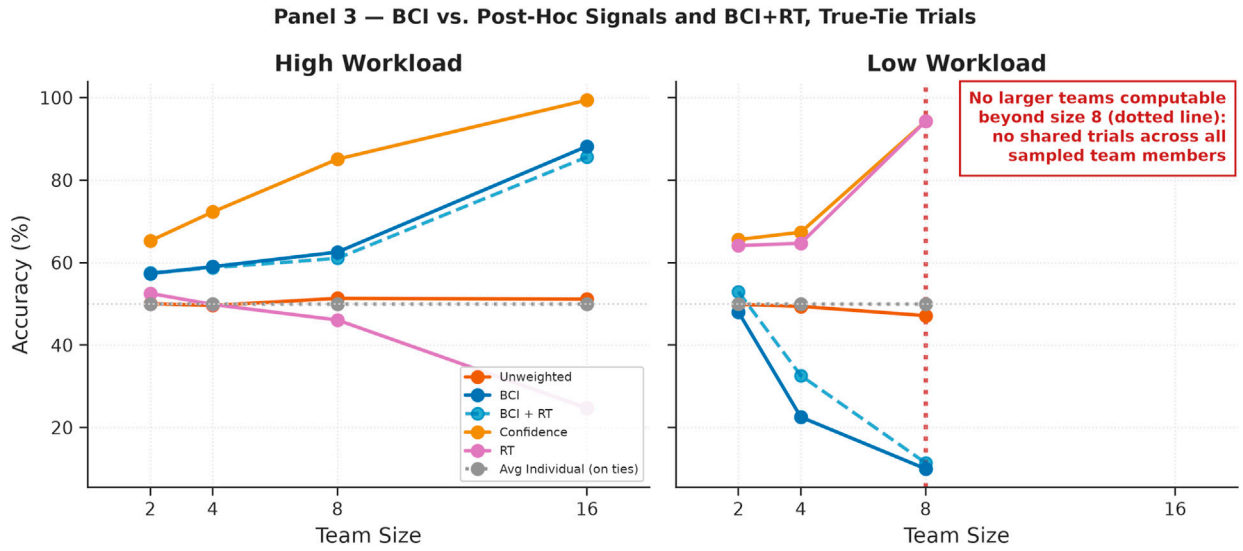

FIGURE 10
Team accuracy on true-tie trials across five aggregation schemes: unweighted vote, pre-emptive neural weighting (BCI), BCI combined with response time, post-hoc confidence weighting, and post-hoc response-time weighting alone, by team size and workload.

true-tie estimate exists at that team size. This context is important for interpreting the team-level gains reported below: while substantial on the trials where they occur, these gains apply to a minority of trials, particularly at larger team sizes, where majority voting alone is already highly accurate (Figure 8).

On these contested trials under High Workload, weighting each vote by the classifier's pre-emptive probability output substantially improved team accuracy, rising from 57.2% at team size 2% to 88.2% at team size 16, while unweighted vote and average individual accuracy remained at approximately 50%, the base rate implied by a tied vote (Figure 9). Under Low Workload, the same weighting scheme produced the opposite effect, collapsing team accuracy toward 10% by team size 8 (Figure 9), consistent with the unreliable, low-true-negative-rate classifier established above: on the one trial subset where its weighting could influence the outcome, it was confidently wrong rather than merely uninformative. To assess the practical, population-level consequence of these tie-restricted gains, we computed accuracy under a selective deployment strategy: applying neural weighting only when a team's raw unweighted vote was tied, and using plain

majority vote otherwise. This is the most conservative strategy available for isolating any potential benefit of the signal. Under High Workload, this strategy produced a consistent, genuine improvement over unweighted majority voting alone, of approximately 0.7 to 2.1 percentage points of overall team accuracy across team sizes 2 to 16. Under Low Workload, by contrast, the same selective strategy produced no improvement, and a small net cost at team sizes 4 and 8 (−1.85 and −0.95 percentage points respectively), confirming that the Low Workload classifier's unreliability is not merely an artefact of the tie-restricted analysis, but persists, and indeed dominates, even under the deployment strategy most conservatively designed to isolate its potential benefit.

Post-hoc behavioural signals are strong but not pre-emptive

For comparison, we evaluated two post-hoc behavioural signals available only after a response is reported: stated confidence and response time (Figure 10). Confidence-weighting was the strongest single signal tested, reaching 99.4% team accuracy at team size 16 under High Workload and 94.3% at team size 8 under Low Workload, in both cases exceeding the pre-emptive neural signal. Response-time weighting, by contrast, underperformed and declined as team size increased. This was not attributable to noise: on tie trials specifically, correct responses were significantly slower than incorrect ones (1.549s vs. 1.466s, $t = 7.43$, $p < .0001$), the reverse of the overall speed-accuracy relationship observed across all trials (Heitz, 2014) (faster responses were more accurate overall, $p = .0045$). Because response-time weighting assumes faster responses are more trustworthy throughout, it is systematically miscalibrated specifically on the trials it is used to adjudicate. Combining the pre-emptive neural signal with response time did not improve on the neural signal alone in either workload condition (Figure 10).

Discussion

The present study set out to test whether spatial-covariance EEG features could provide a genuinely pre-emptive signal of an operator's decision correctness, available before a response is committed, and whether such a signal depends on cognitive workload. We found that it can, but only under High Workload: a Riemannian tangent-space classifier, trained on a strictly motor-safe window, decoded Correct from Incorrect decisions with above-chance, statistically robust accuracy under High Workload, while the equivalent classifier under Low Workload was unreliable, a consequence of near-ceiling behavioural performance leaving too few genuine errors to learn from. At the team level, this signal, used pre-emptively, produced substantial accuracy gains on contested trials specifically under High Workload, while the same weighting scheme actively harmed team accuracy under Low Workload. Together, these findings indicate that a neural decision-reliability signal is not a general-purpose team augmentation tool, but one whose value is conditional on the cognitive state it is deployed under.

The value of a pre-emptive signal

The distinction between pre-emptive and post-hoc evidence, introduced in the Introduction, is not merely conceptual. Post-hoc confidence weighting outperformed the neural signal at the team level in absolute terms, reaching near-ceiling accuracy on contested trials in both workload conditions. This is an important result in its own right: it demonstrates that operators retain meaningful metacognitive access to their own decision reliability even under high workload, consistent with prior work on the dissociability of sensory evidence and behavioural report (Fleming and Lau, 2014; Wokke et al., 2020). However, by definition, confidence cannot inform a decision that has not yet been made, and a team aggregation mechanism that must wait for every member to respond and self-report before weighting their input cannot intervene in time-critical or continuously evolving decisions. The neural signal's comparatively modest team-level gains under High Workload should therefore be read not as a weaker result than confidence weighting, but as a different and complementary one: the only signal tested here that is, in principle, actionable before a response is finalised. Operational systems requiring real-time intervention, rather than post-decision review, have no substitute for this class of signal, regardless of how well post-hoc measures perform once available. A related open question concerns what the classifier is specifically detecting. Because it operates on the full spatial covariance structure of the motor-safe window rather than a fixed set of amplitude-only features, it can in principle capture relational structure across channels beyond raw signal magnitude. However, covariance-based features are not fully independent of amplitude, since larger-amplitude signals will tend to produce larger covariance values, and the grand-average Correct and Incorrect waveforms (Figure 7) differ substantially in amplitude while following broadly similar temporal profiles. We cannot therefore rule out that amplitude differences make a substantial contribution to classifier performance, and we report this as an open question rather than a claim that the classifier necessarily captures pattern information beyond amplitude. A direct comparison against a simpler amplitude-based feature set is a valuable direction for future work.

A further consideration for eventual deployment concerns real-time implementation. The feature window used here is defined retrospectively, ending 100 ms before each trial's own eventual response, which is straightforward in offline analysis but not directly implementable online, where the future response time is not known in advance. We note two directions for addressing this. First, future work could train classifiers per individual rather than pooled across subjects, departing from the leave-one-subject-out validation reported here, which may better capture person-specific neural signatures of decision correctness and support more reliable pre-emptive classification. Second, our own pipeline already demonstrates a reliable, motor-safe, stimulus-locked evoked response at reticle onset (Results, 'The Reticle Elicits a Genuine, Motor-Safe Neural Response'); an equivalent marker generated at the moment a decision-relevant stimulus appears, rather than one anchored to the eventual response, would offer a genuinely prospective and ecologically valid basis for online pre-emptive classification, avoiding the retrospective RT-anchored windowing

used in the present offline analysis. We frame both as directions for future work rather than claims of the present study.

Mechanistic origin and its implications

Ablation analysis indicated that the decoded signal was carried predominantly by posterior occipito-parietal channels (Oz, O2, P3), rather than by central or temporal sites more classically associated with motor preparation (Pfurtscheller and Lopes da Silva, 1999). This is a meaningful methodological safeguard: because features were extracted from a strictly motor-safe window ending 100 ms before each trial's own response, and because the channels driving classification are posterior rather than central, the decoded signal is unlikely to reflect residual motor preparation leaking into the analysis window. Instead, this profile is consistent with the classifier detecting a perceptual or evidence-accumulation process, plausibly related to the neural correlates of error monitoring and metacognitive evaluation reported prior to, or independent of, explicit awareness (Di Gregorio et al., 2020; Schnuerch et al., 2024; Yeung and Summerfield, 2012). We note, however, that this is a correlational, channel-level characterisation rather than a source-localised one, and we return to this as a limitation below.

Trial difficulty and the boundary conditions of the signal

The finding that contested trials, those on which teams could not agree, were substantially explained by the same Target-detection asymmetry established behaviourally (Target trials being intrinsically harder to detect than Non-Target trials to reject) offers a coherent account of when this signal is likely to matter operationally. Because team-level weighting can only influence outcomes on trials where members already disagree, and because such disagreement is concentrated on a specific, identifiable class of trial, the practical value of a pre-emptive neural signal is not diffuse across all decisions but concentrated where teams most need external adjudication. This is a more precise and operationally useful characterisation than a general claim that neural signals "improve team decision-making", and offers a template for identifying, in other domains, which specific decisions a cBCI system is likely to add value to before deployment.

Limitations

Several limitations warrant caution in interpreting and generalising these findings. First, our High/Low Workload labels rest on a task-difficulty manipulation (visual degradation) rather than an independently validated measure of cognitive workload; while manipulations of this kind are an established practice in cBCI and neuroergonomics research (Methods), we did not collect subjective (e.g., NASA-TLX) or physiological corroboration, nor did we assess individual task-relevant skill as a covariate, both of which would be needed to fully dissociate workload from difficulty as constructs. Second, the Low Workload classifier's unreliability, while explained and consistent across independent analyses (raw accuracy, true negative rate, ablation-eligible sample size, and team-level collapse), is itself a consequence of small error-trial counts in that condition, which also limits the number of true-tie trials available for team-level simulation at larger team sizes; a task producing a more even balance of correct and incorrect trials under low workload would be needed to establish whether the absence of a decodable signal under low workload reflects a genuine absence of the underlying neural process, or simply insufficient data to detect it. Third, response-time weighting's failure was traced to a specific, workload- and trial-difficulty-dependent reversal of the speed-accuracy relationship; while this explanation is well-supported by the data reported here, it may not generalise to tasks with different response dynamics, and should not be read as a general claim that response time is an uninformative signal in other contexts. Fourth, team-level results are derived from bootstrap-resampled simulated teams drawn from a single-task, single-session dataset per participant, rather than from teams who trained or worked together; whether these gains persist in genuinely interactive team settings, where members can observe and adapt to one another, remains an open question. Fifth, this revision introduces a substantial number of additional statistical comparisons (signal detection theory criterion and sensitivity, balanced accuracy, ROC-AUC in two conditions, window-duration correlations, and the selective-deployment comparison) beyond those in the original submission; we report Bayes factors alongside p-values throughout partly to mitigate concerns about multiplicity across this expanded set of comparisons, rather than relying on p-values and a formal correction procedure alone. Finally, the task, while implemented in an ecologically motivated continuous virtual reality environment, remains a single perceptual discrimination paradigm; generalisation to other decision domains (e.g., diagnostic, financial) would require independent validation.

Conclusion

EEG-based decoding of decision correctness, using a Riemannian spatial-covariance approach applied to a strictly motor-safe window, provides a genuinely pre-emptive signal of decision reliability that is real, mechanistically grounded in posterior perceptual channels, and translatable into meaningful team-level accuracy gains, but only under high cognitive workload. Under low workload, where behavioural performance is already near ceiling, the same signal and the same aggregation mechanism are unreliable and actively detrimental. This workload-conditionality is not a limitation to be explained away but the central, actionable finding of the study: collaborative brain-computer interface systems should be deployed selectively, informed by the cognitive state of the operators they are intended to support, rather than applied uniformly across all conditions.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://gitlab.com/cmbssd/cbcidm-data (request access as needed).

Ethics statement

The studies involving humans were approved by UK Ministry of Defence Research Ethics Committee (MoDREC, Reference: RQ0000037929). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

CB: Conceptualization, Writing – original draft, Investigation, Visualization, Software, Formal Analysis, Validation, Project administration, Supervision, Data curation, Writing – review and editing, Methodology. SH: Investigation, Writing – review and editing, Data curation. AN: Software, Writing – review and editing, Investigation, Data curation. TR: Funding acquisition, Writing – review and editing, Investigation, Conceptualization, Supervision, Methodology. SF: Funding acquisition, Project administration, Data curation, Conceptualization, Supervision, Writing – review and editing, Methodology, Investigation, Software, Formal Analysis.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was supported by the Defence Science and Technology Laboratory (Dstl) under grant reference (MoDREC, Reference: RQ0000037929).

Acknowledgments

The authors wish to thank the participants for their time and engagement with the experimental protocol.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arshad, S. Z., Zhou, J., Bridon, C., Chen, F., and Wang, Y. (2015). "Investigating user confidence for uncertainty presentation in predictive decision making," in *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, Ozchi '15* (New York, NY, USA: Association for Computing Machinery), 352–360. doi: 10.1145/2838739.2838753

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. (2012). Multiclass brain-computer interface classification by Riemannian geometry. *IEEE Trans. Biomed. Eng.* 59, 920–928. doi: 10.1109/TBME.2011.2172210

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," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Presented at the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Berlin, Germany: IEEE), 3099–3102. doi: 10.1109/EMBC.2019.8856309

Bhattacharyya, S., Valeriani, D., Cinel, C., Citi, L., and Poli, R. (2021). Anytime collaborative brain–computer interfaces for enhancing perceptual group decision-making. *Sci. Rep.* 11, 17008. doi: 10.1038/s41598-021-96434-0

Biernacki, M. P., and Lewkowicz, R. (2024). The role of visual conditions and aircraft type on different aspects of pilot workload. *Appl. Ergon.* 118, 104268. doi: 10.1016/j.apergo.2024.104268

Cabrera, Á. A., Perer, A., and Hong, J. I. (2023). Improving Human-AI collaboration with descriptions of AI behavior. *Proc. ACM Hum-Comput Interact.* 7 (136), 1–136. doi: 10.1145/3579612

Canonico, L. B., Flathmann, C., and McNeese, N. (2019). Collectively intelligent teams: integrating team cognition, collective intelligence, and AI for future teaming. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 63, 1466–1470. doi: 10.1177/1071181319631278

Cinel, C., Fernandez-Vargas, J., Tremmel, C., Citi, L., and Poli, R. (2022). Enhancing performance with multisensory cues in a realistic target discrimination task. *PLOS ONE* 17, e0272320. doi: 10.1371/journal.pone.0272320

Congedo, M., Barachant, A., and Bhatia, R. (2017). Riemannian geometry for EEG-Based brain-computer interfaces; a primer and a review. *Brain-Comput. Interfaces* 4, 155–174. doi: 10.1080/2326263X.2017.1297192

Dehais, F., Ladouce, S., Darmet, L., Nong, T.-V., Ferraro, G., Torre Tresols, J., et al. (2022). Dual passive reactive brain-computer interface: a novel approach to human-machine symbiosis. *Front. Neuroergonomics* 3, 824780. doi: 10.3389/fnrgo.2022.824780

Di Gregorio, F., Maier, M. E., and Steinhauser, M. (2020). Are errors detected before they occur? Early error sensations revealed by metacognitive judgments on the timing of error awareness. *Conscious. Cogn.* 77, 102857. doi: 10.1016/j.concog.2019.102857

Dirican, A. C., and Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Comput. Sci., World Conf. Inf. Technol.* 3, 1361–1367. doi: 10.1016/j.procs.2011.01.016

Ablin, P., Cardoso, J.-F., and Gramfort, A. (2018). Faster independent component analysis by preconditioning with Hessian approximations. *IEEE Trans. Signal Process.* 66, 4040–4049. doi: 10.1109/TSP.2018.2844203

Fleming, S. M., and Lau, H. C. (2014). How to measure metacognition. *Front. Hum. Neurosci.* 8, 443. doi: 10.3389/fnhum.2014.00443

Galton, F. (1907). Vox populi. *Nature* 75, 450–451. doi: 10.1038/075450a0

Goldfarb, A., and Lindsay, J. R. (2022). Prediction and judgment: why artificial intelligence increases the importance of humans in war. *Int. Secur.* 46, 7–50. doi: 10.1162/isec_a_00425

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* 7, 267. doi: 10.3389/fnins.2013.00267

Hamada, D., Nakayama, M., and Saiki, J. (2020). Wisdom of crowds and collective decision-making in a survival situation with complex information integration. *Cogn. Res. Princ. Implic.* 5, 48. doi: 10.1186/s41235-020-00248-z

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

Hassan, J., Reza, Md.S., Ahmed, S. U., Anik, N. H., and Khan, Md.O. (2024). EEG workload estimation and classification: a systematic review. *J. Neural Eng.* 22. doi: 10.1088/1741-2552/ad705e

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd. *Behav. Res. Methods Instrum. Comput.* 27, 46–51. doi: 10.3758/BF03203619

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Front. Neurosci.* 8, 150. doi: 10.3389/fnins.2014.00150

Kerr, N. L., and Tindale, R. S. (2004). Group performance and decision making. *Annu. Rev. Psychol.* 55, 623–655. doi: 10.1146/annurev.psych.55.090902.142009

Koopman, C., and Zammit-Mangion, D. (2024). Artificial intelligence for real-time tolerance to critical flight data errors in large aircraft. *J. Aerosp. Inf. Syst.* 21, 726–734. doi: 10.2514/1.I011365

Kosch, T., Karolus, J., Zagermann, J., Reiterer, H., Schmidt, A., and Woźniak, P. W. (2023). A survey on measuring cognitive workload in human-computer interaction. *ACM Comput. Surv.* 55 (283), 1–283. doi: 10.1145/3582272

Kothe, C., Shirazi, S. Y., Stenner, T., Medine, D., Boulay, C., Grivich, M. I., et al. (2024). The lab streaming layer for synchronized multimodal recording. *bioRxiv* 2024, 13.580071. doi: 10.1101/2024.02.13.580071

Ledoit, O., and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* 88, 365–411. doi: 10.1016/S0047-259X(03)00096-4

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-Data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

McKinney, W. (2010). "Data structures for statistical computing in python," in *Presented at the Python in Science Conference*. Texas: Austin, 56–61. doi: 10.25080/Majora-92bf1922-00a

Murayama, K., Blake, A. B., Kerr, T., and Castel, A. D. (2016). When enough is not enough: information overload and metacognitive decisions to stop studying information. *J. Exp. Psychol. Learn. Mem. Cogn.* 42, 914–924. doi: 10.1037/xlm0000213

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pfurtscheller, G., and Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.* 110, 1842–1857. doi: 10.1016/S1388-2457(99)00141-8

Plainis, S., and Murray, I. J. (2002). Reaction times as an index of visual conspicuity when driving at night. *Ophthalmic Physiol. Opt. J. Br. Coll. Ophthalmic Opt. Optom.* 22, 409–415. doi: 10.1046/j.1475-1313.2002.00076.x

Poli, R., Valeriani, D., and Cinel, C. (2014). Collaborative brain-computer interface for aiding decision-making. *PLoS ONE* 9, e102693. doi: 10.1371/journal.pone.0102693

Quitadamo, L. R., Cavrini, F., Sbernini, L., Riillo, F., Bianchi, L., Seri, S., et al. (2017). Support vector machines to detect physiological patterns for EEG and EMG-Based human–computer interaction: a review. *J. Neural Eng.* 14, 011001. doi: 10.1088/1741-2552/14/1/011001

Sadras, N., Sani, O. G., Ahmadipour, P., and Shanechi, M. M. (2023). Post-stimulus encoding of decision confidence in EEG: toward a brain–computer interface for decision making. *J. Neural Eng.* 20, 056012. doi: 10.1088/1741-2552/acec14

Salvatore, C., Valeriani, D., Piccialli, V., and Bianchi, L. (2022). Optimized collaborative brain-computer interfaces for enhancing face recognition. *IEEE Trans. Neural Syst. Rehabil. Eng.* 30, 1223–1232. doi: 10.1109/TNSRE.2022.3173079

Schnuerch, R., Schmuck, J., and Gibbons, H. (2024). Cortical oscillations and event-related brain potentials during the preparation and execution of deceptive behavior. *Psychophysiology* 61, e14695. doi: 10.1111/psyp.14695

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. doi: 10.1038/s41598-017-08265-7

Valeriani, D., Cinel, C., and Poli, R. (2019). Brain–computer interfaces for human augmentation. *Brain Sci.* 9, 22. doi: 10.3390/brainsci9020022

Valeriani, D., O'Flynn, L. C., Worthley, A., Sichani, A. H., and Simonyan, K. (2022). Multimodal collaborative brain-computer interfaces aid human-machine team decision-making in a pandemic scenario. *J. Neural Eng.* 19, 056036. doi: 10.1088/1741-2552/ac96a5

Wickens, C. D. (2008). Multiple resources and mental workload. *Hum. Factors* 50, 449–455. doi: 10.1518/001872008X288394

Wokke, M. E., Cleeremans, A., and Ridderinkhof, K. R. (2017). Sure i'm sure: prefrontal oscillations support metacognitive monitoring of decision making. *J. Neurosci.* 37, 781–789. doi: 10.1523/JNEUROSCI.1612-16.2016

Wokke, M. E., Achoui, D., and Cleeremans, A. (2020). Action information contributes to metacognitive decision-making. *Sci. Rep.* 10, 3632. doi: 10.1038/s41598-020-60382-y

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science* 330, 686–688. doi: 10.1126/science.1193147

Yeung, N., and Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 367, 1310–1321. doi: 10.1098/rstb.2011.0416

Yger, P., Spampinato, G. L., Esposito, E., Lefebvre, B., Deny, S., Gardella, C., et al. (2018). A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings *in vitro* and *in vivo*. *eLife* 7, e34518. doi: 10.7554/eLife.34518

Yu, K., Prasad, I., Mir, H., Thakor, N., and Al-Nashash, H. (2015). Cognitive workload modulation through degraded visual stimuli: a single-trial EEG study. *J. Neural Eng.* 12, 046020. doi: 10.1088/1741-2560/12/4/046020

Zhou, Y., Huang, S., Xu, Z., Wang, P., Wu, X., and Zhang, D. (2022). Cognitive workload recognition using EEG signals and machine learning: a review. *IEEE Trans. Cogn. Dev. Syst.* 14, 799–818. doi: 10.1109/TCDS.2021.3090217