Human-AI Teaming Under Deception: An Implicit BCI Safeguards Drone Team Performance in Virtual Reality

Christopher Baker
c.baker@qub.ac.uk
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast

Stephen Hinton
s.f.hinton@ljmu.ac.uk
School of Psychology
Liverpool John Moores University

Akashdeep Nijjar
an22851@essex.ac.uk
School of Computer Science and Electronic Engineering
University of Essex

Riccardo Poli
rpoli@essex.ac.uk
School of Computer Science and Electronic Engineering
University of Essex

Caterina Cinel
ccinel@essex.ac.uk
School of Computer Science and Electronic Engineering
University of Essex

Tom Reed
treed@mail.dstl.gov.uk
Defence Science Technology Laboratory

Stephen Fairclough
s.fairclough@ljmu.ac.uk
School of Psychology
Liverpool John Moores University

Core Narrative: The paper argues that under high-stakes deception, a purely **Neuro-Decoupled Team (NDT)** strategy, based only on EEG-SVM Confidence, is unusually robust. The traditional reliance on behavioural (RT/Subjective Confidence) and external AI data collapses, while the BCI signal maintains its integrity, achieving resilience through a strategic neural shift.

Abstract

Human-AI teams can be vulnerable to catastrophic failure when feedback from the AI is incorrect, especially under high cognitive workload. Traditional team aggregation methods, such as voting, are susceptible to these AI errors, which can actively bias the behaviour of each individual and inflate the likelihood of an erroneous group decision. We hypothesised that a collaborative Brain-Computer Interface (cBCI) using neural activity collected before a behavioural decision is made can provide a source of information that is "decoupled" from this biased behaviour, thereby protecting the team from the deleterious influence of AI error. We tested this in a VR drone surveillance task where teams of operators faced high workload and systematically misleading AI cues, comparing traditional behaviour-based team strategies against a purely Neuro-Decoupled Team (NDT) that used only BCI confidence scores derived from pre-response EEG. Under AI deception, behaviour-based teams catastrophically failed, with Majority Vote accuracy collapsing to 44%. The NDT, however, maintained 98% accuracy, a statistically significant synergistic gain over even the team's best individual performer (p < .001). This was explained by a neuro-behavioural decoupling, where the BCI's predictions remained highly accurate while the operator's subjective confidence became an unreliable signal. We conclude that an implicit BCI provides resilience by learning to adapt its neural strategy, shifting from relying on signals of efficient, 'autopilot' processing in simple conditions to interpreting signatures of effortful deliberation when confronted with cognitive conflict. This demonstrates a system that leverages the context of the neural signal to defend against AI-induced error in high-stakes environments.

Introduction

Modern human-AI teaming holds immense promise for enhancing operational effectiveness across a range of important decision domains, from everyday recommendations to critical workplace predictions in fields like medicine, law, or financial services [1,2]. This has led to a drive to improve collaboration by developing better mental models of AI behaviour [3] and calibrating human trust in the AI system [4–6]. However, this promise is tempered by the fact that humans often struggle to rely on AI appropriately, leading to sub-optimal team performance [7]. This introduces novel vulnerabilities, particularly when a systematically flawed or deceptive AI provides guidance that induces correlated errors across multiple human operators [1,4]. Under conditions of high cognitive demand [8,9], this risk is amplified. The statistical advantage of a group [10,11], the "wisdom of crowds" is predicated on the independence of its members' judgments. A flawed AI, acting as a common source of biased information, can systematically violate this independence [12]. This can invert the statistical advantage of a team, transforming the group into a mechanism for catastrophic, widespread failure [13,14].

Conventional team aggregation methods are designed to harness this collective intelligence by condensing individual inputs into a single team output [15,16]. Simple strategies like majority voting treat all members equally, while more sophisticated approaches apply weightings to individual inputs based on metrics like past performance or self-reported confidence [17,18]. Both approaches, however, are particularly susceptible to the failure mode introduced by a deceptive AI because they rely on the implicit assumption that the behavioural outputs of team members, their decisions and their confidence, are reliable indicators of ground truth. This assumption is fundamentally challenged by contemporary models of metacognition (see [19] for review), which frame subjective confidence not as a direct readout of decision evidence, but as a separate, "second-order" computation that makes an inference about the likely quality of a decision [20,21]. Empirical work supports this theoretical separation, demonstrating that the sensory evidence contributing to a first-order decision can be dissociated from the information that supports a subsequent metacognitive judgment [21]. This confidence-accuracy relationship is especially fragile under conditions of high cognitive

demand or information overload, where flawed metacognitive assessments can lead to paradoxically sub-optimal performance [22]. A misleading AI can therefore systematically corrupt this distinct metacognitive channel, creating a state where operators become confidently incorrect. The existence of distinct neural processes for performance monitoring, which can even give rise to "early error sensations" before a motor response is executed [23], further highlights that the decision and its evaluation are separable processes and thus independently vulnerable to bias. This invalidation of behavioural confidence as a reliable signal creates a critical gap in current safeguards, as the aggregation of corrupted behavioural reports can transform the team into a mechanism for widespread failure.

As a potential safeguard, the collaborative Brain-Computer Interface (cBCI) offers access to an insulated channel of evidence. A cBCI can be designed to bypass the potentially corrupted conscious cognitive process by accessing implicit neural activity generated before a final behavioural judgment is formed. This pre-decisional neural signal, it is argued, can remain insulated from the cognitive biases that affect the overt response, providing a more robust source of information under deception [24,25]. The neuroscientific plausibility for this approach is grounded in the brain's generation of reliable neural markers that differentiate between effortful, deliberative processing and efficient, automatic states [26]. When individuals are confronted with conflicting information that requires active problem-solving, a specific neural signature reliably emerges from frontal midline brain regions: an increase in theta-band (~4–8 Hz) power. This signal is considered a general marker for the need to engage cognitive control across many different types of cognitive challenges [27], reflecting a genuine neural oscillation rather than a simple evoked response [28], and is believed to originate from a dedicated neural microcircuit for conflict detection [29]. This same conflict-monitoring system also gives rise to the brain's automatic and rapid responses to errors and negative feedback, known respectively as the Error-Related Negativity (ERN) and Feedback-Related Negativity (FRN) [30]. Crucially, the conscious feeling of having made an error can even arise before a physical response is executed, supporting the existence of a pre-decisional signal of internal conflict [23]. Following this initial conflict detection, a more sustained and effortful re-evaluation of evidence is indexed by a late positive potential (LPP), an event-related potential associated with successful cognitive reappraisal and conscious deliberation [31]. Conversely, when tasks can be performed automatically and without conflict, the brain enters a more efficient 'autopilot' state. This state is characterized by its own distinct set of 'honest signals'. One key signature is the modulation of posterior alpha rhythms (8-13 Hz), where increased power is a well-established marker of reduced visual engagement and the active inhibition of irrelevant sensory information, a process known as attentional gating [32,33]. A second, key signature is the modulation of sensorimotor beta power (13-30 Hz). Beta oscillations are characteristically suppressed during active movement but are prominent during steady-state motor control, reflecting the maintenance of the current motor 'status quo' [34]. More than just a signal of motor idling, this beta activity has been shown to index the brain's confidence in its internal models; higher beta power reflects a higher confidence in the current motor plan and a reduced need for adaptation [35]. From a computational perspective, this corresponds to a lower weighting of sensory prediction errors, effectively promoting stability [36] and response certainty [36,37].

The objective of the current study is therefore twofold: first, to test the hypothesis that a purely neuro-decoupled team (NDT), which relies exclusively on implicit BCI data, can provide resilience against AI-induced deception; and second, to move beyond a simple performance demonstration to reveal the adaptive neural strategy that enables this resilience. To achieve this, we compare the performance of an NDT against traditional behaviour-based teams within a high-workload drone surveillance task featuring a deceptive AI. The core analysis focuses on the pre-decisional ReticleOn epoch to isolate the predictive neural signals that underpin the BCI's success via the offline team simulations.

Results

The following results demonstrate the unique resilience of a purely neuro-decoupled team (NDT) under high-stakes deception. To isolate this effect, the analysis focuses exclusively on the critical experimental condition where teams operated under high cognitive workload and were presented with systematically incorrect AI guidance. This "deception condition" is designed to model a worst-case scenario where traditional, behaviour-based team strategies are most vulnerable to the correlated error induced by a misleading AI. The findings are presented in three stages: first, we establish the NDT's macro-level resilience at the system level; second, we demonstrate the micro-level neuro-behavioural decoupling that explains its success; and third, we identify the specific neurophysiological mechanism driving the BCI's adaptive strategy.

Context: Team Performance in High-Demand Scenarios

Overall Team Performance (High Workload, ReticleOn)

Under High Workload conditions, considering all trials (Fig. 1), various team aggregation methods were compared. The standard Majority Human method achieved accuracies ranging from 84.35% (N=2) to 89.65% (N=8). behavioural weighting methods included RT Weighted Human (accuracies: 85.60% for N=2 to 89.34% for N=8) and Subjective Confidence Weighted Human (accuracies: 87.14% for N=2 to 90.77% for N=8). The combined human-only method, RT + Subjective Confidence Human, resulted in accuracies from 87.56% (N=2) to 90.30% (N=8). The BCI-only aggregation, SVM Confidence Weighted BCI, produced accuracies in the range of 76.15% (N=2) to 90.87% (N=8).

The mixed methods, integrating human and BCI information, generally demonstrated the strongest overall performance. The Subjective Confidence + SVM Confidence Mixed method showed high accuracies, increasing from 90.04% (N=2) to 93.58% (N=4), 95.01% (N=6), and reaching 95.96% for eight-person teams. The comprehensive RT + Subjective Confidence + SVM Confidence Mixed method also achieved high accuracies, from 90.46% (N=2) to 93.70% (N=4), 94.86% (N=6), and 95.53% (N=8). The RT + SVM Confidence Mixed method yielded accuracies from 89.51% (N=2) to 94.97% (N=8).

Notably, for larger team sizes, these mixed methods surpassed the average accuracy of the Best Individual Average (which was 91.27% for N=6 and 92.22% for N=8). For N=8 teams, the Subjective Confidence + SVM Confidence Mixed method achieved 95.96%, the RT + Subjective Confidence + SVM Confidence Mixed method achieved 95.53%, and the RT + SVM Confidence Mixed method achieved 94.97%, all exceeding the Best Individual Average of 92.22%. This suggests a synergistic benefit. The Average Individual Average was consistently around 84.45%.
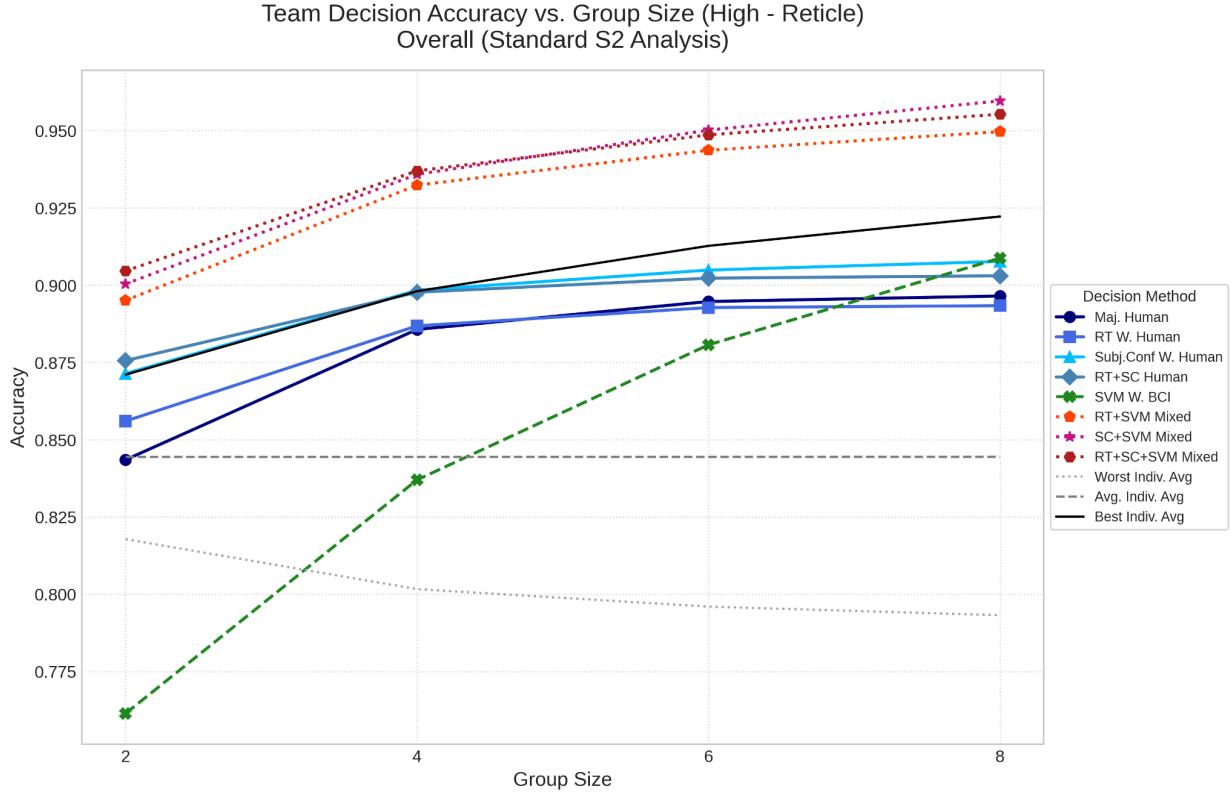
Fig. 1. Team Decision Accuracy vs. Group Size - Overall (High - Reticle)

Team Performance with Correct In-Task AI (High Workload, ReticleOn)

When the in-task AI provided a correct cue (Fig. 2), team accuracies for all methods generally showed substantial improvement, often approaching near-perfect performance, particularly for larger team sizes. Human-only methods like Majority Human achieved accuracies from 91.74% (N=2) to 99.39% (N=8). The mixed methods incorporating BCI information demonstrated exceptionally high performance under these favorable AI cue conditions. For instance, at a group size of N=8, the Subjective Confidence + SVM Confidence Mixed method reached an accuracy of 99.61%, and the RT + Subjective Confidence + SVM Confidence Mixed method achieved 99.78%. These accuracies significantly exceeded the performance of the Best Individual Average under these AI-correct conditions, which was 92.23% for N=8, highlighting a strong synergistic effect when both human and AI inputs were reliable.
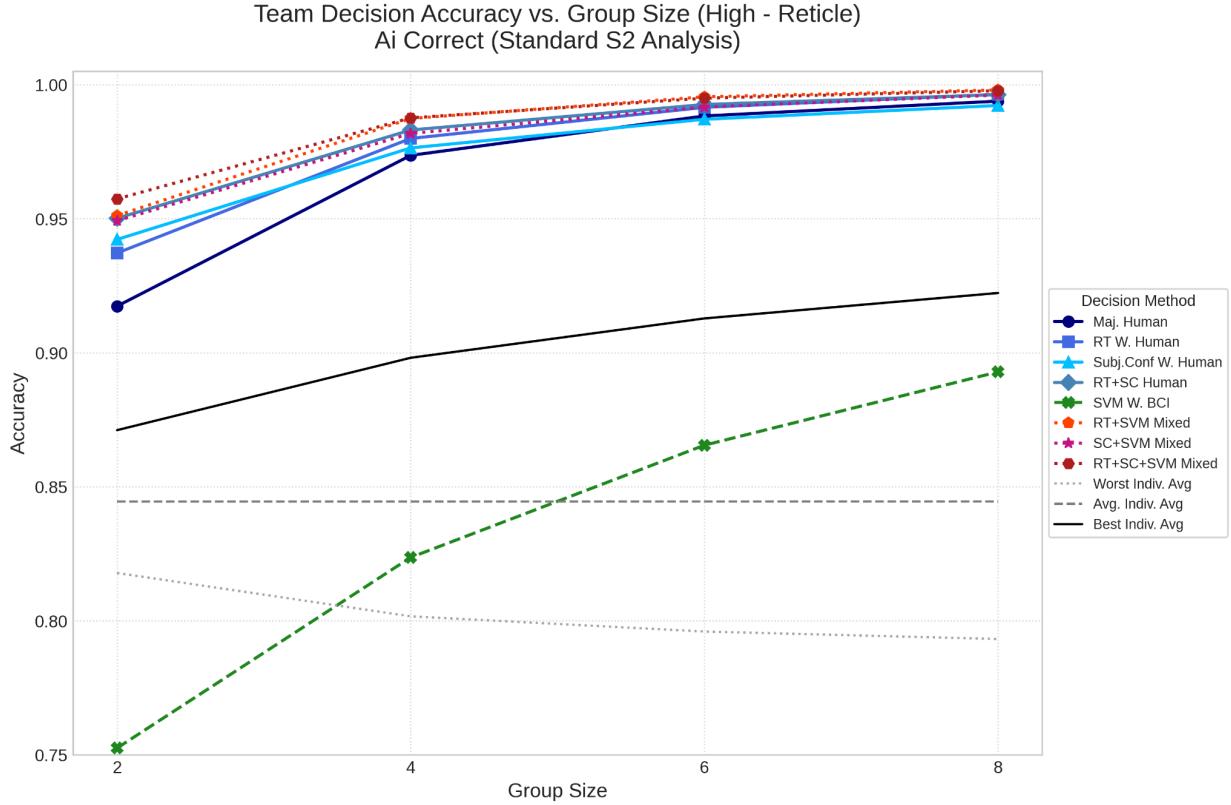
Fig. 2. Team Decision Accuracy vs. Group Size (High - Reticle) - AI Correct

Team Performance with Incorrect In-Task AI (High Workload, ReticleOn)

Conversely, when the in-task AI cue was incorrect (Fig. 3), overall team performance was substantially lower across all methods compared to when the AI was correct, as expected. The Majority Human method yielded accuracies from 44.41% (N=2) to 43.95% (N=8), indicating a strong negative influence of the misleading AI on simple human aggregation. However, mixed methods still often outperformed simple human aggregation, demonstrating a capacity to mitigate some of the AI's negative impact. For example, at N=8, the Subjective Confidence + SVM Confidence Mixed method achieved 78.84%, and the RT + Subjective Confidence + SVM Confidence Mixed method achieved 75.65%. Even under these challenging AI-incorrect conditions, the best mixed methods demonstrated accuracies often superior to the Majority Human and approached, though did not consistently surpass, the Average Individual Average (which was 84.43% for N=8 on these trials). The Best Individual Average for these specific AI-incorrect trials was 92.17% for N=8, which the team methods did not reach, underscoring the difficulty of overcoming systematically incorrect AI guidance.
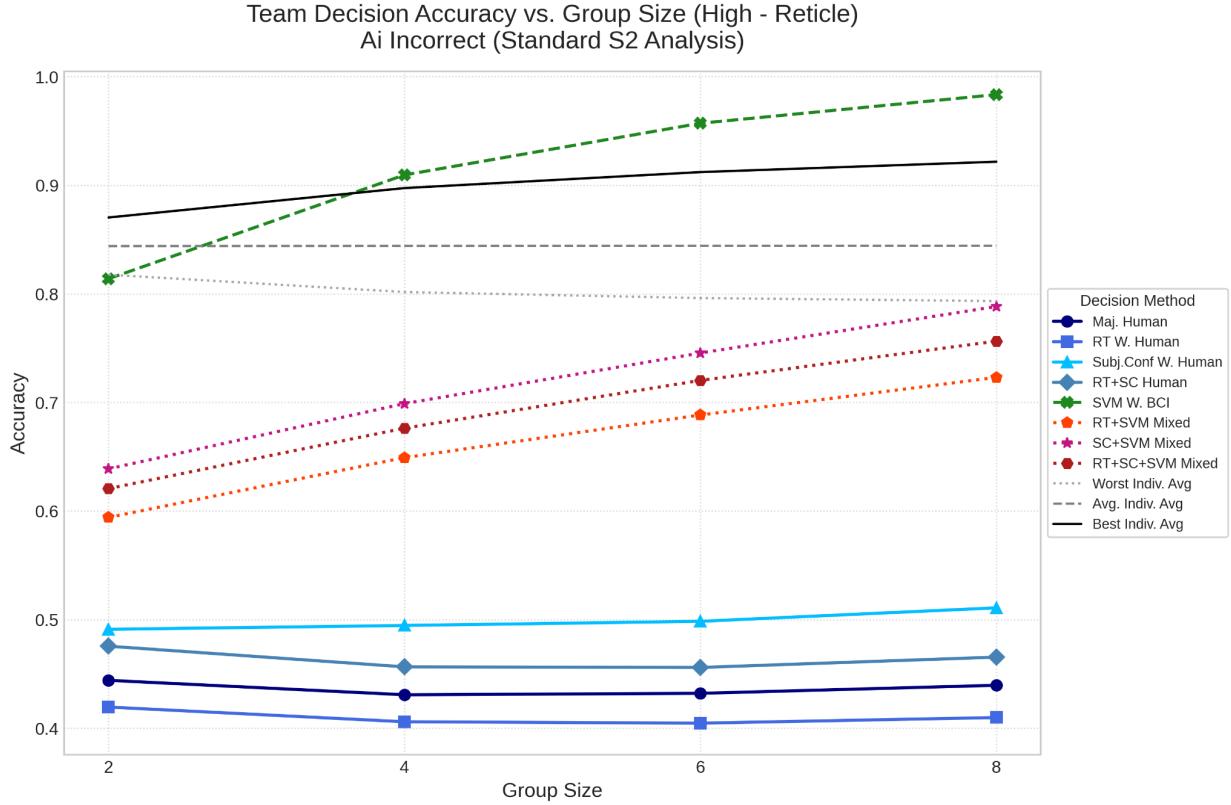
Fig. 3. Team Decision Accuracy vs. Group Size (High - Reticle) - AI Incorrect

Macro-Level Resilience: The Neuro-Decoupled Team Outperforms Failing Behavioural Methods

When confronted with systematically incorrect AI guidance under high workload, teams relying on traditional behavioural aggregation methods demonstrated a catastrophic failure, with the Majority Vote method collapsing to 44% accuracy and Subjective Confidence faring marginally better at 51%. In stark contrast, the purely Neuro-Decoupled Team (NDT), which aggregated decisions using only BCI-SVM confidence scores, maintained a profound resilience by achieving 98% accuracy. This core finding, see Figure 4, 5, Table. 1, represents a true synergistic gain, as the NDT's performance significantly surpassed not only the average individual (84%) but also the mean accuracy of the team's best individual performer (92%; $p < .001$). This confirms that under conditions of high-stakes deception, the implicit neural signal provides a more reliable source for group decision-making than the operators' explicit behavioural outputs. This demonstrates that even at a team size of N=8, the correlated error induced by the AI is sufficient to overwhelm the statistical advantage typically gained from group aggregation. Furthermore, paired t-tests on the outcomes of over 7 million simulated team decisions confirmed the NDT's superiority to both the Majority Vote ($t = 598.75$, $p < .001$) and Subjective Confidence methods ($t = 518.30$, $p < .001$).
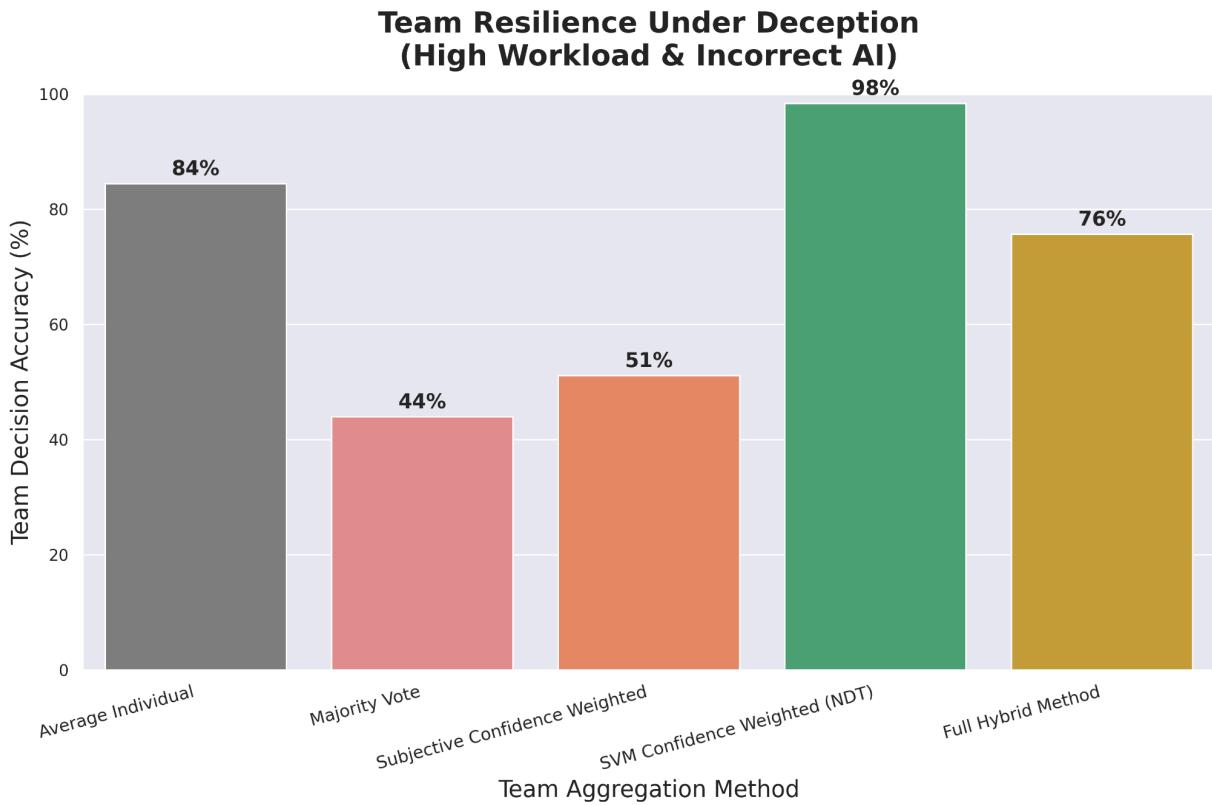
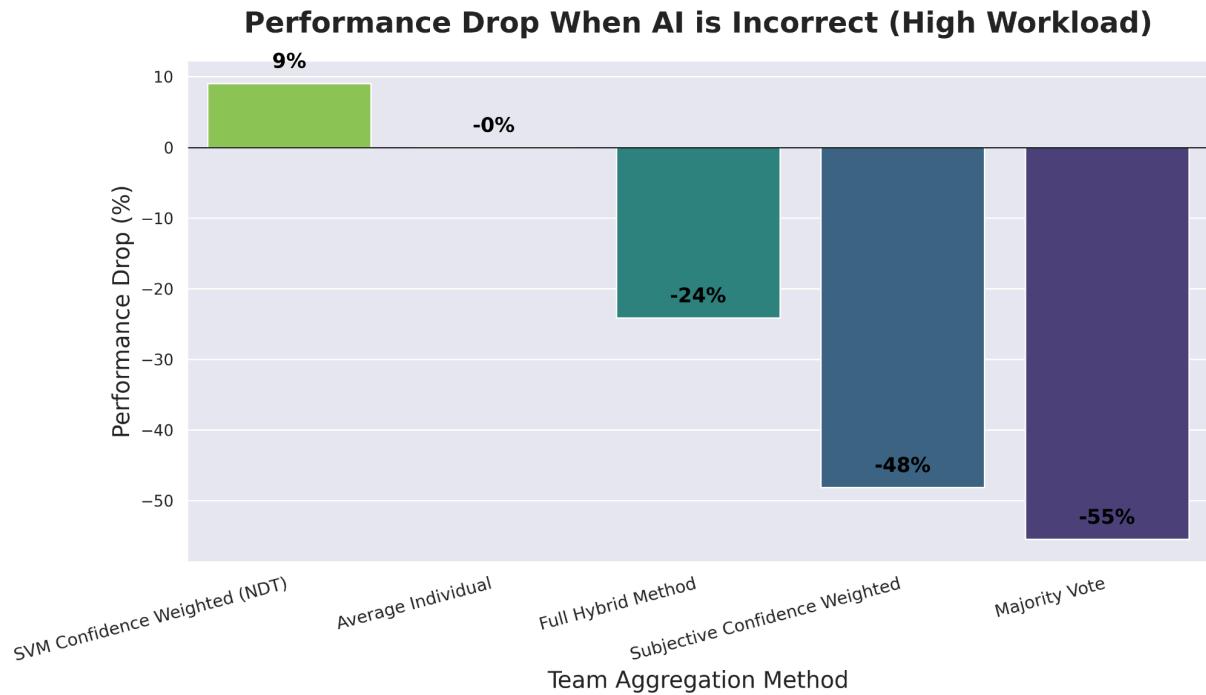Fig. 4. Team n=8 Resilience: High Workload Condition - AI Incorrect

Fig. 5. Team n=8 Performance Drop - AI Incorrect

Group Size	Method	Accuracy AI Incorrect	Accuracy Drop
8	SVM Confidence Weighted (NDT)	0.9835	9.06
8	Average Individual	0.8443	-0.02
8	Full Hybrid Method	0.7565	-24.13
8	Subjective Confidence Weighted	0.5109	-48.13
8	Majority Vote	0.4395	-55.44

Table. 1. Team n=8 Performance Drop - AI Incorrect

Micro-Level Proof: The Decoupling of Neural and behavioural Fidelity

The NDT's resilience is explained by a fundamental decoupling between the fidelity of the implicit neural signal and the explicit behavioural report under deception. A correlation analysis revealed that during these high-workload, incorrect AI trials, the relationship between an operator's subjective confidence and the actual ground truth correctness was markedly weak, though statistically significant ($r = .194$, $p < .001$). In stark contrast, the linear correlation between the BCI's SVM-derived confidence score and ground truth was effectively non-existent ($r = -.017$, $p = .380$). This neuro-behavioural decoupling, illustrated in Figure 6, demonstrates that the BCI preserved its predictive integrity by accessing a neural substrate insulated from the cognitive biases that corrupted the operator's conscious sense of certainty. While the SVM's confidence output does not show a simple linear relationship with correctness, the underlying neural patterns it classifies remain a profoundly reliable source of evidence, allowing the classifier's predictions to achieve the high fidelity that underpins the NDT's success. To formally test this decoupling, we used Steiger's Z-test for comparing dependent correlations, as both metrics shared a common variable (ground truth correctness) from the same sample. The test confirmed that the difference in these correlation coefficients was statistically significant ($Z = 7.80$, $p < .001$), providing quantitative evidence for the neuro-behavioural decoupling effect.
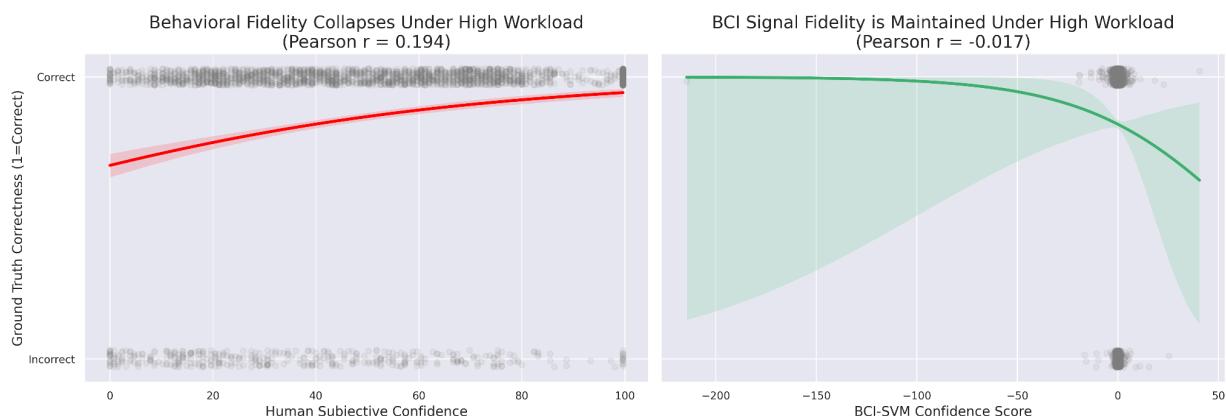

Fig. 6. Correlation Plots of Neuro-behavioural Decoupling of Signal Fidelity under Deception Human Decisions vs. BCI Classifier

Neurophysiological Mechanism: The BCI's Two Decision Strategies

A comparison of the top-performing features in each workload condition reveals that the BCI learns to employ two distinct and functionally relevant strategies. Table 2 lists the top 10 most influential EEG features when the classifier was operating under High Workload, while Table 3 lists the top 10 for the Low Workload condition. An analysis of these lists, visually summarised in Figure 7, demonstrates a clear strategic adaptation.

In the Low Workload condition, the BCI learns a "Rhythmic Stability" strategy. The most predictive features are dominated by measures of neural consistency and efficiency. Critically, the single most important feature is low variance at F9 (Var_F9), indicating that the *absence* of executive effort in the prefrontal cortex is the strongest sign of a correct, automatic decision. This is supported by the other top features, which are primarily alpha and beta power modulations over centro-parietal areas, a signature of a brain on "autopilot," efficiently gating sensory information.

In contrast, under the deceptive High Workload condition, this stability-based strategy fails. The BCI adaptively shifts to an "Effort & Instability" strategy. It discovers that the most reliable predictors are now signals of active cognitive struggle. The top features become the maximum amplitude at executive sites (Max_F9, Max_FC6) and high variance over the motor cortex (Var_C4), reflecting moments of intense cognitive work and response uncertainty. This demonstrates that the BCI's resilience comes from its ability to recognise when its primary "autopilot" strategy is unreliable and to switch to a secondary strategy that successfully interprets the neural signatures of active, effortful deliberation.
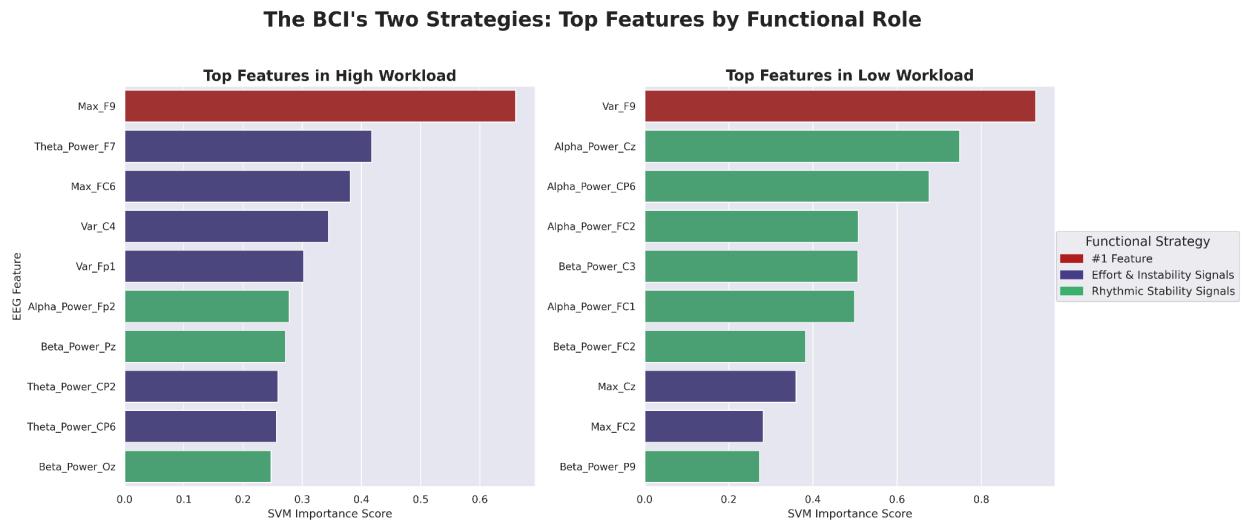
Fig. 7. Feature importance for AI Incorrect HW vs. LW

Feature	Importance High Workload	Importance Low Workload	Importance Delta
Max_F9	0.6606	0.1508	0.5098
Theta_Power_F7	0.4176	0	0.4176
Max_FC6	0.3812	0	0.3812
Var_C4	0.3443	0.1633	0.181
Var_Fp1	0.3025	0	0.3025
Alpha_Power_Fp2	0.27795	0.1025	0.17545
Beta_Power_Pz	0.2717	0	0.2717

Theta_Power_CP2	0.2592	0.0142	0.245
Theta_Power_CP6	0.25665	0	0.25665
Beta_Power_Oz	0.2475	0	0.2475

Table. 2. Top 10 Features HW

Feature	Importance Low Workload	Importance High Workload	Importance Delta
Var_F9	0.9293	0	-0.9293
Alpha_Power_Cz	0.749	0.2308	-0.5182
Alpha_Power_CP6	0.6755	0	-0.6755
Alpha_Power_FC2	0.5076	0	-0.5076
Beta_Power_C3	0.5071	0	-0.5071
Alpha_Power_FC1	0.4987	0	-0.4987
Beta_Power_FC2	0.38235	0.17375	-0.2086
Max_Cz	0.3599	0	-0.3599
Max_FC2	0.2817	0.1192	-0.1625
Beta_Power_P9	0.2733	0	-0.2733

Table. 3. Top 10 Features LW

Further neurophysiological evidence for these two distinct decision-making strategies is provided by the grand-averaged event-related potentials (ERPs), time-locked to the stimulus onset (see Figure 8). A direct comparison of the brain's electrical response in each condition reveals a profound difference in the underlying neural processing. In the Low Workload condition, the neural response is characteristic of efficient, stereotyped processing. Following a sharp negative deflection around 350ms, the activity in all channel groups rapidly returns toward the baseline. This pattern signifies a quick, automatic decision that does not require prolonged, effortful cognitive engagement, consistent with the BCI's "Rhythmic Stability" strategy. In stark contrast, the High Workload condition reveals a pattern of intense, sustained deliberation. While the initial negative-going wave is still present, it is immediately followed by a large, sustained positive-going wave, a late positive potential (LPP), that is most prominent in the frontal channel groups. This LPP is a well-established marker of extended, controlled processing, cognitive re-evaluation, and the allocation of attentional resources to resolve conflict, providing a clear neural correlate for the BCI's "Effort & Instability" strategy.

The pivotal role of the F9 site is clearly illustrated by these ERPs. In the Low Workload condition, the F9 waveform is highly consistent, explaining why low signal variance (Var_F9) was the most powerful predictor of a correct "autopilot" decision. Conversely, in the High Workload condition, the F9 trace exhibits a dramatic deflection from its negative peak to a sustained positive peak. This large amplitude swing is precisely what the classifier learned to interpret as a signal of active, effortful deliberation, making maximum amplitude (Max_F9) the most important feature for its "effort" strategy. Therefore, the ERPs provide clear temporal evidence that the BCI's resilience is driven by its ability to track two functionally and neurophysiologically distinct modes of decision-making.
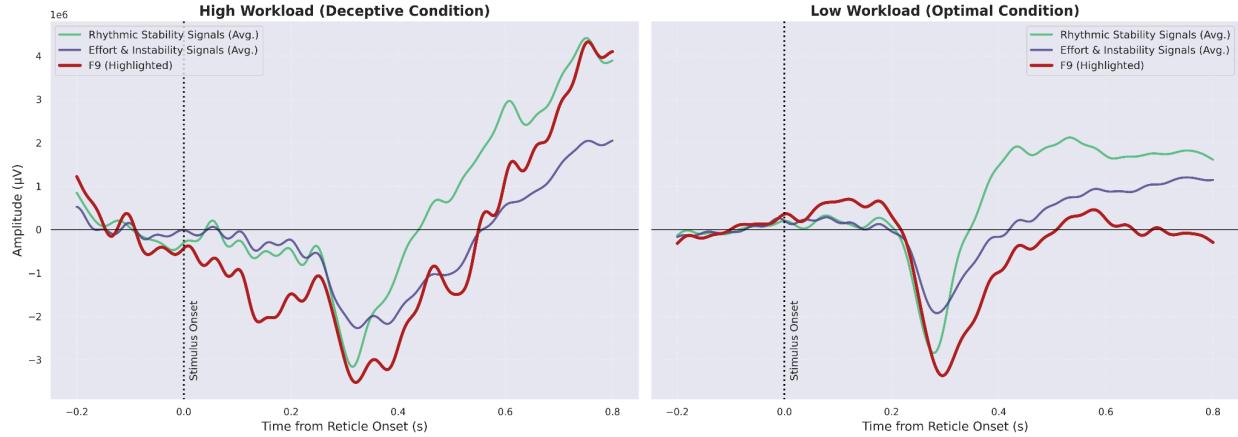
Fig. 8. Feature importance for AI Incorrect HW vs. LW

Discussion

The central finding of this study is that a purely neuro-decoupled team (NDT), relying solely on an implicit BCI strategy, demonstrates profound resilience to systematic AI deception in a high-workload environment. While traditional team aggregation methods based on behavioural outputs like voting or subjective confidence collapsed to near chance-level performance, the NDT maintained 98% accuracy. This synergistic gain, significantly outperforming even the team's best individual, confirms our primary hypothesis: pre-decisional neural activity provides a channel of evidence that is uniquely insulated from the cognitive biases induced by a misleading external cue. The catastrophic failure of behavioural teams demonstrates that under conditions of correlated error, the wisdom of the crowd can be inverted, with larger groups becoming more susceptible to mass failure. The NDT's success proves that an implicit BCI can serve as a powerful neuroergonomic safeguard against this critical vulnerability in human-AI systems.

The Two Strategies: An Adaptive Neural Mechanism for Resilience

Our analysis revealed that the BCI's resilience is not based on a single, static neural marker, but on an adaptive, two-strategy system. In simple, low-workload conditions, the classifier learned a "Rhythmic Stability" strategy, primarily leveraging the absence of executive effort (low variance at F9) and the clean, predictable alpha/beta modulations of a brain on "autopilot." This is an efficient, low-cost strategy for predictable environments. However, when confronted with the deceptive AI cue in the high-workload condition, this "autopilot" signal was shattered. Here, the BCI adaptively switched to an "Effort & Instability" strategy, discovering that the most reliable predictors of a correct decision were now the neural signatures of active, effortful deliberation. Critically, the BCI was able to distinguish between these two strategies using neural data from the ReticleOn epoch alone, well before the operator initiated a motor response. This pre-decisional decoding confirms that the system is tapping into the early, implicit process of deliberation itself, not merely the certainty of the final action, which is key to its insulating properties. The ERP results provide converging temporal evidence for this switch, showing a massive and sustained late positive potential (LPP) emerging only in the high-workload condition, a clear hallmark of extended, controlled processing. This demonstrates a sophisticated form of machine learning, where the BCI has not just learned to classify a pattern, but has learned to identify the context in which its primary strategy is failing and to switch to a secondary, more robust strategy.

Implications for Human-AI Teaming and Neuroergonomics

The findings have significant implications for the design of future human-AI systems, particularly in high-stakes operational environments. First, they serve as a stark warning against an over-reliance on explicit behavioural metrics like confidence ratings, which this study shows can become systematically corrupted and decoupled from ground truth under deception. An operator can be behaviourally confident yet neurophysiologically conflicted.

Second, this study provides a blueprint for a new class of neuroergonomic safety nets. A real-time system implementing this BCI could act as an "implicit dissent channel," flagging team decisions where there is a dangerous mismatch between high behavioural consensus and high underlying neural conflict. Such a system could prompt a team to "think again," preventing catastrophic errors in domains like remote piloting, intelligence analysis, or cybersecurity, where a single misleading automated cue could have severe consequences.

Limitations and Future Work

Several limitations must be acknowledged in the context of results presented. The primary limitation is that the team performance was evaluated via offline simulation. The next crucial step is to implement this NDT strategy in a real-time, closed-loop collaborative BCI to study live team interactions and validate the synergistic gains. Secondly, our deception paradigm was standardised. A valuable follow-up study could directly test the BCI's resilience against different AI feedback mechanisms within the high-workload state, for example, by contrasting an AI with consistent errors against one with intermittent or varied performance. Furthermore, it must be acknowledged that in a real-world drone scenario, operators would likely have full control of the vehicle. This active piloting introduces additional cognitive demands and agency, which may mitigate or increase the effect of AI errors in taxing scenarios, an interaction that warrants future investigation. Building on the pivotal role of the F9 site revealed in our analysis, a dedicated follow-up study could seek to quantify whether specific F9-related features (e.g., max amplitude) serve as the key neural indicator of successful intervention against AI error. Such a study could test if this F9 response is a distinguishing electrophysiological feature of this specific human-AI interaction effect, potentially providing a real-time biomarker for when an operator is actively overcoming deceptive guidance. Finally, while the participant-specific SVMs proved effective, future research should explore more operationally viable generalised classifiers. As part of this, the predictive quality of our pre-stimulus event period could be further studied to enhance the BCI's potential, possibly enabling even earlier detection of likely error states.

Conclusion

This study demonstrates that an implicit BCI strategy, leveraging the brain's automatic conflict-monitoring systems, provides a robust and effective defense against the correlated errors induced by deceptive AI. By learning to distinguish between neural signatures of effortless "autopilot" and effortful deliberation, the BCI was able to maintain near-perfect predictive accuracy while traditional, behaviour-based teams catastrophically failed. This work moves beyond simply decoding decisions, showcasing a system that interprets the *context* of the neural signal to achieve resilience. It provides a clear path toward developing safer and more effective human-AI partnerships, where the implicit wisdom of the brain can be harnessed to protect teams in high-stakes, uncertain environments.

Methods

Participants

Seventeen participants were included in data analyses ([10 Female]; Mean age ± SD = [24.2 ± 5.04). An initial cohort of 20 individuals was recruited; however, data from three participants were excluded prior to the main analysis due to criteria established in preliminary quality control checks, including insufficient data quality after EEG preprocessing, or significant deviations in trial sequence alignment during task execution due to technical issues and the need for alignment of trials across all combinations of team group size. All included participants reported normal or corrected-to-normal vision and no history of neurological disorders or particular susceptibility to VR-induced motion sickness. Prior to the experiment, participants provided written informed consent and completed the Barratt Impulsiveness Scale (BIS-11) [38] and the Balloon Analogue Risk Task (BART) prior to commencing the study. Participants received monetary compensation for their participation. The experimental protocol received favourable opinion by UK MoDREC, App No: 2309/MODREC/24 Ref: RQ0000037929 and all procedures were conducted in accordance with the ethical standards outlined in the Declaration of Helsinki.

Study Design

The study employed a within-subject repeated measures design to evaluate the effectiveness of a collaborative Brain-Computer Interface (cBCI) system designed to enhance group decision-making in a dynamic virtual reality (VR) environment. The primary within-subject factor manipulated was cognitive workload, presented at two levels (Low vs. High) across distinct experimental blocks. Electroencephalographic (EEG) data were analysed time-locked to two different events: the onset of a targeting reticle ('ReticleOn'), representing a pre-stimulus anticipatory period, and the participant's response ('ButtonPress'), capturing peri- and post-decisional neural activity. Team performance was assessed through offline simulations, evaluating decision accuracy for simulated groups of varying sizes (2, 4, 6, and 8 members) under different decision aggregation algorithms. Individual behavioural metrics, including accuracy, response time (RT), and subjective confidence ratings, served as primary dependent variables at the individual level. Additionally, personality traits related to impulsivity and risk-taking were assessed for exploratory analysis. While EEG data were analyzed time-locked to both 'ReticleOn' and the participant's response ('ButtonPress'), the results were highly convergent. As our primary hypothesis concerned the insulating properties of pre-decisional neural activity, all results reported herein are based exclusively on the 'ReticleOn' epoch analysis.
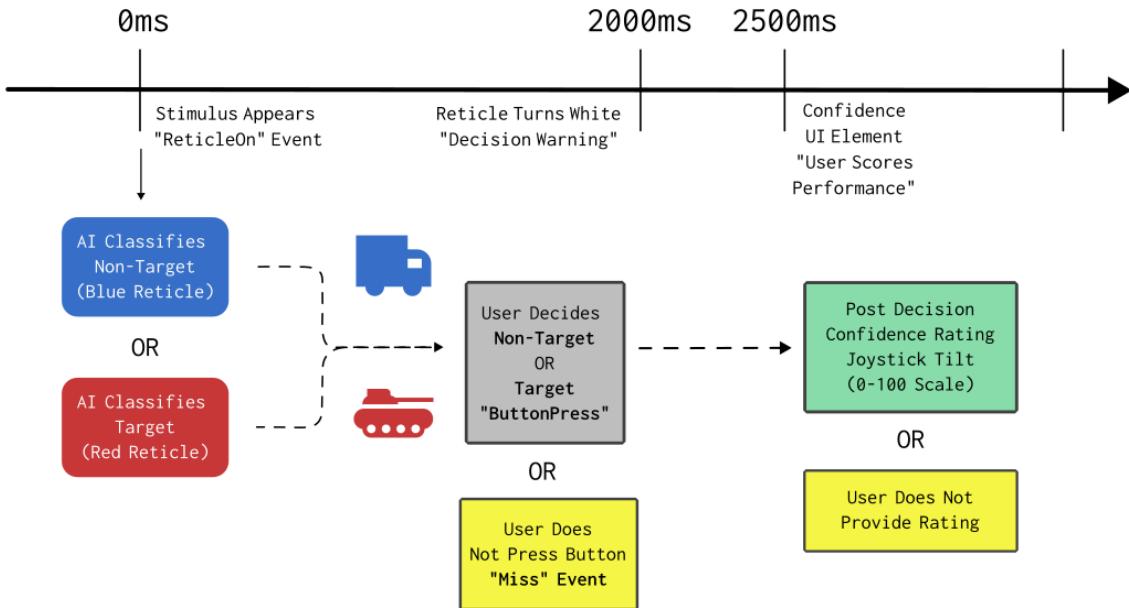
BCI Trial Time (Single Trial Logic)

Fig. 9 Timeline of single trial within task

VR Drone Task and Procedure

Participants were seated in the laboratory while they viewed the virtual environment via a Varjo Aero HMD. The simulation, developed in Unreal Engine 5 and rendered on a high-performance PC, depicted the viewpoint of a quadcopter type drone flying over a simulated landscape designed to be semi realistic i.e.textured realistically for a temperate climate and populated with sparse foliage but tailored to not interfere with stimuli acquisition (See Fig. 10). Task stimuli consisted of 3D models appearing on the ground: Non-Targets (See Fig. 11) and Targets (See Fig. 12). Participants completed six 8-minute blocks, with each block containing 50 trials (30 Non-Targets, 20 Targets). Stimuli were evenly spaced 300m apart, but varied in lateral placement (-30m, -15m, 0m, +15m, +30m)] and rotation (0°, 90°, 180°, 270°).

Fig. 10. View of the virtual landscape from the simulated drone (Low Workload Condition)

Fig. 11. Example of a Non-target stimulus

Fig. 12. Example of a Target stimulus

Each trial followed a consistent sequence. Initially, the drone flew to a set virtual height (25m) and advanced across the landscape. As stimuli came into view a targeting reticle appeared over the stimuli ('ReticleOn'). Crucially, the colour of this reticle immediately indicated an in-task AI's assessment of the stimulus: the reticle appeared blue if the AI classified the stimulus as a Non-Target (see Fig. 13), and red if the AI classified it as a Target (see Fig. 14). Participants were informed that this AI-indicated reticle colour was intended to assist their decision-making but that the AI was not perfect and they should make their own final judgment. The AI-indicated reticle remained on-screen for a total of 2500ms. Participants were tasked with discriminating the stimulus type and responding as quickly and accurately as possible by pressing the designated joystick button for Target or Non-Target. Participants were instructed to respond whilst the reticle was on-screen. To provide a response deadline warning, 2000ms after the 'ReticleOn' event, the reticle would change to white for 500ms and then disappear. (See Fig. 15). Any responses made after the reticle disappeared were not counted, and those trials were classed as misses. Following this primary response ('ButtonPress'), a prompt appeared, requiring participants to rate their confidence in the preceding decision via tilting the joystick to select a value on a 0-100 scale (0 = Not Confident, 100 = Very Confident) (See Fig. 16).
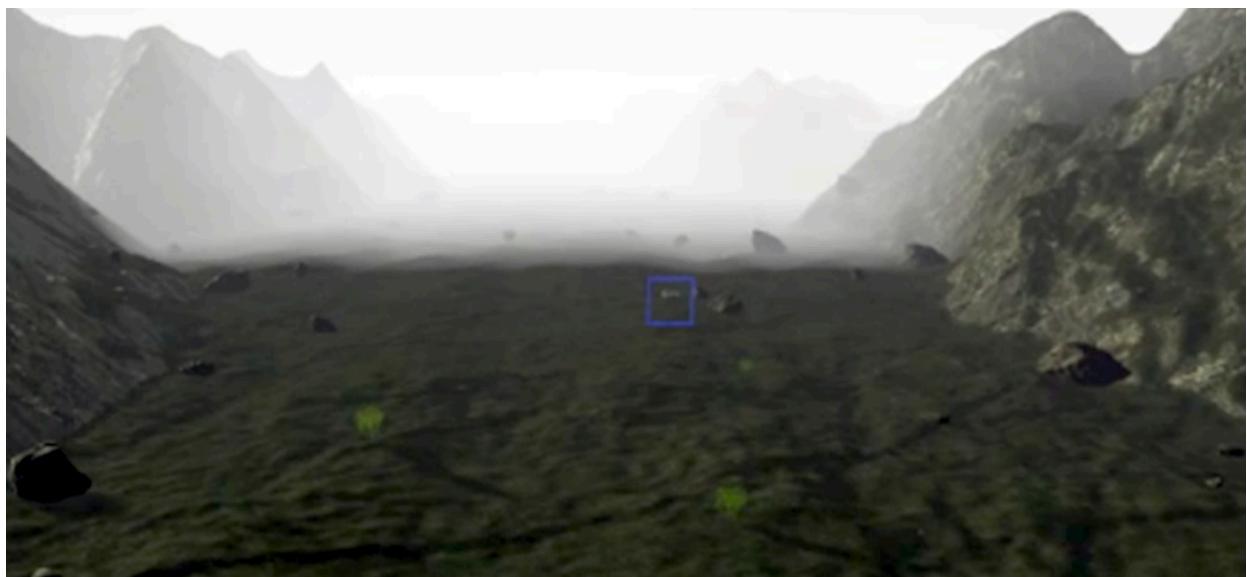
Fig. 13. Reticle appears blue indicating non-target feedback from AI.

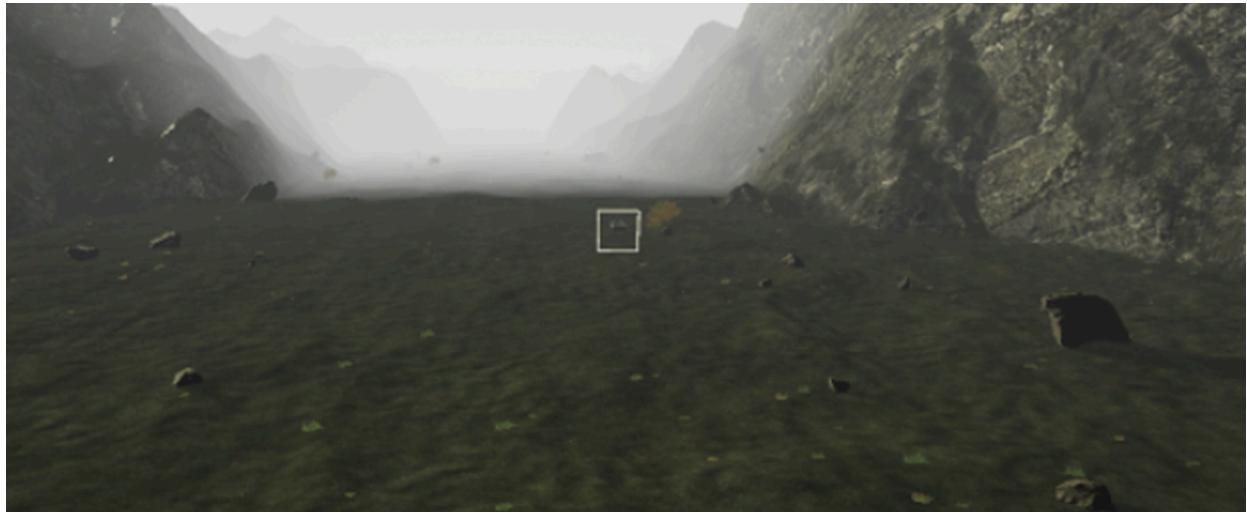Fig. 14. Reticle appears red indicating target feedback from AI.

Fig. 15 Reticle response warning

Fig. 16. User interface element for subjective confidence rating

Cognitive workload was manipulated across different blocks of trials. The Low Workload condition had the virtual light level fixed as approximate for midday average daylight. Conversely, the High Workload condition reduced the light level by 50%, and the solar angle by 9° simulating a night condition (See Fig. 18). Participants first completed a practice block for each workload condition, followed by three blocks under each workload condition.

Fig. 18. View from the simulated drone in the High Workload (low light) condition

Data Acquisition

Continuous EEG data were recorded using a 32-channel LiveAmp system (Brain Products GmbH). Electrodes were arranged according to the international 10-20 configuration using an electrode cap (actiCAP snap electrode cap). Average referencing was utilised, the actiCAP utilised a dedicated ground electrode point at the front and centre of the head, between Fp1 and Fp2. Electrode impedances were kept below 30KΩ throughout the recording. The EEG signal was recorded at a sampling rate of 500Hz. behavioural data, including response button presses (for RT calculation) and subsequent confidence ratings, were logged via joystick actions which generated LabStreamingLayer[39] (LSL) markers. Event markers corresponding to critical task events (e.g., ReticleOn, StimulusOn, ButtonPress) were generated by the Unreal Engine environment using the LSL UE5 plugin and transmitted via LSL. Both EEG and event marker streams were simultaneously recorded and synchronised using LabRecorder and OpenSignals software, resulting in .xdf files for each session.

Procedure

Upon arrival at the laboratory, participants were briefed on the study's objectives and procedures. After providing written informed consent, they completed the BART and BIS-11 questionnaires. Participants were then fitted with the EEG cap and the VR HMD, see Fig. 19. Electrode impedances were checked and adjusted to be below.

After baselining, participants undertook a training task of two blocks (one for each workload condition) then the VR drone task. They received instructions on the target/non-target discrimination, the response mapping (button presses), and the confidence rating scale. The main experiment consisted of six blocks, divided equally between the Low Workload and High Workload conditions (three blocks per condition). The entire experimental session, including setup, task execution, and debriefing, lasted approximately 150 minutes per participant. Upon completion, participants were debriefed and received their monetary compensation.

Fig. 19. Participant partaking in the study

EEG Signal Processing

EEG data were processed offline using MNE-Python [40] and custom Python scripts. The processing pipeline included:

1. Loading and Filtering: Data from each XDF file were loaded. A band-pass filter (0.1–30 Hz, FIR) and a 50 Hz notch filter were applied to the data.
2. Trimming: The continuous recording was trimmed to the duration of the experimental task using the first and last LSL markers.
3. Artifact Rejection (ICA): Independent Component Analysis (ICA) was employed to identify and remove stereotypical artifacts such as eye blinks and lateral eye movements. To enhance ICA quality, data from all sessions for a given participant were concatenated. FastICA was run on this concatenated data. Components reflecting ocular or other non-neural artifacts were manually identified by visual inspection of their topography and time course and were flagged for removal. The corresponding ICA demixing matrix was then applied to the individual session files to project out the artifactual components.
4. Epoching: The cleaned continuous data were segmented into epochs relative to two primary events: (i) 'ReticleOn' epochs (-200 ms to +800 ms relative to marker onset) and (ii) 'ButtonPress' epochs (-500 ms to +300 ms relative to marker onset).
5. Baseline Correction: Epochs were baseline-corrected using the pre-event interval: -200 ms to 0 ms for ReticleOn epochs, and -500 ms to -200 ms for ButtonPress epochs.
6. Trial Validation: Specific trials identified as problematic during preliminary checks (e.g., Trial 18) were excluded during the creation of the final metadata associated with the epochs.

BCI Feature Extraction and Classification

A participant-specific BCI was developed using an SVM classifier to predict the likelihood of decision correctness based on EEG features.

Feature Extraction: For each epoch (ReticleOn or ButtonPress), features were extracted per channel (32 channels). Time-domain features included the mean amplitude, maximum amplitude, and variance across the epoch. Frequency-domain features included the average Power Spectral Density (PSD) within the Theta (4–8 Hz), Alpha (8–13 Hz), and Beta (13–30 Hz) bands, estimated using Welch's method (1-40 Hz range).

SVM Training: For each participant, an SVM classifier (sklearn.svm.SVC[41]) was trained. Feature vectors were first standardised (StandardScaler). SelectKBest with mutual_info_classif scoring identified the top 5 features. To handle potential class imbalance (more correct than incorrect trials), the feature-selected training data was balanced using RandomUnderSampler. Hyperparameters (kernel type: 'rbf'/'linear'; C: 0.1-100; gamma: 'scale'/'auto') were optimised via 5-fold stratified cross-validation using GridSearchCV on this balanced training set, maximising accuracy.

Prediction and Confidence Score: The optimised SVM for each participant was then applied to predict the label (1=Correct Prediction, 0=Incorrect Prediction) for all of their valid, feature-selected (but unbalanced) trials. The classifier's decision function output for each trial, representing the signed distance from the separating hyperplane, was recorded as the SVM Confidence score. This score served as a BCI-derived measure of decision certainty.

Team Simulation Procedure and Aggregation Methods

Performance of simulated teams was evaluated offline using an exhaustive combinatorial approach. First, the per-trial data from all participants was compiled into a single dataset. This dataset was filtered to remove trials with response times over 2.6 seconds or trials with SVM confidence scores that were statistical outliers (beyond 1.5 × interquartile range). From this cleaned dataset, three key metrics were normalized globally on a 0-1 scale across all trials and participants: (i) Response Time, where faster responses received higher scores (1 - normalized value); (ii) Subjective Confidence, where ratings were scaled from 0-100 to 0-1; and (iii) BCI-SVM Confidence, where the absolute value of the SVM's decision function output was min-max scaled. These normalized values served as decision weights in the subsequent simulations.

For each of the approximately 150 valid experimental trials within each workload condition (Low and High), team performance was simulated independently for every possible unique combination of participants drawn from the final N=17 cohort for each specified team size (m = 2, 4, 6, and 8).

This meant that for any single trial, decisions were simulated for:

- all 136 unique two-person teams (the number of distinct combinations of 2 participants from 17),
- all 2,380 unique four-person teams (the number of distinct combinations of 4 participants from 17),
- all 12,376 unique six-person teams (the number of distinct combinations of 6 participants from 17),

- all 24,310 unique eight-person teams (the number of distinct combinations of 8 participants from 17).

Given approximately 150 trials per workload condition, and two workload conditions (Low and High, totalling ~300 trials available for analysis after individual trial validation), the total number of simulated team decisions generated was substantial. For two-person teams, this involved 136 unique team combinations, each assessed over approximately 300 trials, resulting in approximately 40,800 simulated decisions. For four-person teams, this involved 2,380 unique team combinations, resulting in approximately 714,000 simulated decisions. For six-person teams, this involved 12,376 unique team combinations, resulting in approximately 3,712,800 simulated decisions. For eight-person teams, this involved 24,310 unique team combinations, resulting in approximately 7,293,000 simulated decisions. In total, this comprehensive simulation strategy yielded over 11.7 million individual team decision points for analysis.

For these per-trial, per-team-composition simulations, data were grouped by unique trial identifiers ensuring that for each simulated team decision, only data (e.g., individual response, RT, SVM confidence) from participants who had experienced that identical trial were included when constituting that specific team instance. Team decisions were aggregated using several methods, detailed below and summarized in Table 4. For all weighted methods, ties were resolved by favouring the 'Target' classification.

1. Majority Human: Each member's behavioural response (Target/Non-Target) contributed one vote. The team decision was the most frequent response, with ties broken randomly.
2. RT Weighted Human: Each member's behavioural response was weighted by their Normalised_RT. The team decision was the response type with the higher sum of weights.
3. Subjective Confidence Weighted Human: Each member's behavioural response was weighted by their Confidence_Rating_Norm.
4. RT + Subjective Confidence Human: Each member's behavioural response was weighted by the average of their Normalised_RT and Confidence_Rating_Norm.
5. SVM Confidence Weighted BCI (NDT): This purely neuro-decoupled method used the BCI's predicted label for each member, weighted by the Normalised_SVM_Confidence.
6. RT + SVM Confidence Mixed: This hybrid method integrated behavioural and neural data. For each member, evidence for a decision was calculated as a 50/50 split between their RT-weighted behavioural response and their SVM-confidence-weighted BCI prediction. The team decision was based on the highest total summed evidence.
7. Subjective Confidence + SVM Confidence Mixed: A hybrid method similar to the above, but using the Confidence_Rating_Norm for the behavioural component.
8. Full Hybrid (RT + Subjective Confidence + SVM Confidence Mixed): The most comprehensive method. For each member, a "human component" score was calculated from the average of their Normalised_RT and Confidence_Rating_Norm. The final evidence was then a 50/50 split between this human component score (applied to the behavioural response) and the BCI component score (applied to the BCI prediction).

Individual Performance Baselines: For robust comparison, team performance was benchmarked against the average accuracy of the best, worst, and average human performer within each specific simulated group on each trial.

Table 4: Calculation Summary of Key Team Decision Methods

Method Label in Plot	Core Logic Description	Key Information Used per Team Member (Trial-Level)
Worst Individual Avg	Average of the lowest individual accuracy within each unique simulated team.	Overall Human Accuracy (of the worst performer)
Average Individual Avg	Average of the mean individual accuracy within each unique simulated team.	Overall Human Accuracy (mean of individuals)
Best Individual Avg	Average of the highest individual accuracy within each unique simulated team.	Overall Human Accuracy (of the best performer)
Majority Human	Each member's response contributes one vote. Team decision is the most frequent.	Human Response (Target/Non-Target)
RT Weighted Human	Human responses weighted by Normalised_RT	Human Response, Normalised RT
Subj. Conf Weighted Human	Human responses weighted by Confidence_Rating_Norm	Human Response, Confidence_Rating_Norm
RT + Subj. Conf Human	Human responses weighted by the average of Normalised_RT and Confidence_Rating_Norm.	Human Response, Normalised_RT, Confidence_Rating_Norm
SVM Conf Weighted BCI	SVM-predicted labels weighted by Normalised_SVM_Confidence.	SVM Predicted Label, Normalised_SVM_Confidence
RT + SVM Conf Mixed	Evidence is 0.5 * (Human Score from RT) + 0.5 * (BCI Score from SVM Conf).	Human Response, Normalised_RT, SVM Predicted Label, Normalised_SVM_Confidence
Subj. Conf + SVM Conf Mixed	Evidence is 0.5 * (Human Score from Subj. Conf) + 0.5 * (BCI Score from SVM Conf).	Human Response, Confidence_Rating_Norm, SVM Predicted Label, Normalised_SVM_Confidence
RT + Subj. Conf + SVM Conf Mixed	Human score = avg(Normalised_RT, Confidence_Rating_Norm). Evidence is 0.5 * (Human Score) + 0.5 * (BCI Score from SVM Conf).	Human Response, Normalised_RT, Confidence_Rating_Norm, SVM Predicted Label, Normalised_SVM_Confidence

Statistical Analysis

Individual behavioural data (accuracy, RT, confidence) were analysed to assess the impact of the Workload manipulation (Low vs. High). Depending on data distributions, paired t-tests or Wilcoxon signed-rank tests were used for continuous variables (RT, confidence), while Chi-square tests were used for accuracy (comparing counts of correct/incorrect decisions). For simulated team performance, mean accuracies for the proposed cBCI weighting method(s) were compared against baseline methods (Majority, Best/Average Individual) for each group size using paired t-tests or Wilcoxon tests. Corrections for multiple comparisons (e.g., Bonferroni) were applied where appropriate. Statistical significance was defined at an alpha level of $p < 0.05$. All statistical analyses were performed using Python [42] and its scientific computing libraries, primarily SciPy, for significance testing. Data processing and manipulation were conducted using Pandas [43] and NumPy [44]. Visualisations were generated with Matplotlib[45] and Seaborn [46].

Hypotheses

Based on the premise that pre-decisional neural activity provides an insulated channel of evidence against external deception, we formulated three primary hypotheses:

H1: Resilience of the Neuro-Decoupled Team
We predicted that under conditions of high workload and misleading AI guidance, a team decision strategy based solely on implicit BCI data (the NDT) would yield significantly higher accuracy than strategies based on explicit behavioural data (e.g., Majority Vote). We further hypothesised that this neuro-decoupled approach would produce a synergistic gain, with the NDT's accuracy surpassing that of the team's best individual performer.

H2: Neuro-behavioural Decoupling
We predicted that under the influence of deceptive AI cues, the predictive utility of operators' explicit behavioural reports would degrade. Specifically, we hypothesised that the correlation between subjective confidence and ground truth correctness would weaken, while the BCI classifier, drawing from an insulated neural channel, would maintain its ability to accurately predict decision outcomes.

H3: Context-Dependent Neural Predictors
We hypothesised that the neural features most predictive of a correct decision would be fundamentally different depending on the task context. Based on established neurocognitive literature, we predicted:

- (a) In the Low Workload condition, characterised by simple, automatic processing, the most predictive features would be associated with efficient attentional gating (e.g., modulations of posterior alpha and beta rhythms).
- (b) In the High Workload (deceptive) condition, which demands active cognitive control to overcome conflict, the most predictive features would shift to markers of executive engagement and conflict monitoring (e.g., increased mid-frontal theta power).

Declarations

Ethics approval and consent to participate

The experimental protocol received favourable opinion by the UK Ministry of Defence Research Ethics Committee (MoDREC), Application Number: 2309/MODREC/24, Reference: RQ0000037929. All procedures were conducted in accordance with the ethical standards outlined in the Declaration of Helsinki. Written informed consent was obtained from all individual participants included in the study.

Data Availability Statement

The datasets generated and/or analysed during the current study are not publicly available due to restrictions imposed by the funding body (Defence Science and Technology Laboratory - Dstl). However, data are available from the corresponding author (CB) on reasonable request and subject to a data sharing agreement, if appropriate and in accordance with Dstl policy.

Competing interests

The authors declare that they have no competing interests.

Funding

Authors' contributions

C.B.: Conceptualisation, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing.
S.H.: Conceptualisation, Methodology, Formal Analysis, Investigation, Writing – Review & Editing.
S.F.: Conceptualisation, Methodology, Software, Formal Analysis, Investigation, Writing – Review & Editing.
A.N.: Conceptualisation, Methodology, Writing – Review & Editing.
R.P.: Conceptualisation, Methodology, Writing – Review & Editing.
C.C.: Conceptualisation, Methodology, Writing – Review & Editing.
T.R.: Writing – Review & Editing.

References

1. Erlei, A., Sharma, A. & Gadiraju, U. Understanding Choice Independence and Error Types in Human-AI Collaboration. in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* 1–19 (Association for Computing Machinery, New York, NY, USA, 2024). doi:10.1145/3613904.3641946.

2. Goldfarb, A. & Lindsay, J. R. Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War. *Int. Secur.* **46**, 7–50 (2022).

3. Cabrera, Á. A., Perer, A. & Hong, J. I. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proc ACM Hum-Comput Interact* **7**, 136:1-136:21 (2023).

4. Li, J. *et al.* Understanding the Effects of Miscalibrated AI Confidence on User Trust, Reliance, and Decision Efficacy. Preprint at https://doi.org/10.48550/arXiv.2402.07632 (2025).

5. Roeder, L. *et al.* A Quantum Model of Trust Calibration in Human–AI Interactions. *Entropy* **25**, 1362 (2023).

6. Canonico, L. B., Flathmann, C. & McNeese, N. Collectively Intelligent Teams: Integrating Team Cognition, Collective Intelligence, and AI for Future Teaming. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **63**, 1466–1470 (2019).

7.   He, G., Buijsman, S. & Gadiraju, U. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proc ACM Hum-Comput Interact* **7**, 276:1-276:29 (2023).

8.   Koopman, C. & Zammit-Mangion, D. Artificial Intelligence for Real-Time Tolerance to Critical Flight Data Errors in Large Aircraft. *J. Aerosp. Inf. Syst.* **21**, 726–734 (2024).

9.   Possibilities and Challenges for Artificial Intelligence in Military Applications.

10. GALTON, F. Vox Populi. *Nature* **75**, 450–451 (1907).

11. Kerr, N. L. & Tindale, R. S. Group Performance and Decision Making. *Annu. Rev. Psychol.* **55**, 623–655 (2004).

12. Orzechowski, K. P., Sienkiewicz, J., Fronczak, A. & Fronczak, P. When the crowd gets it wrong – the limits of collective wisdom in machine learning. *Sci. Rep.* **15**, 22139 (2025).

13. Jayles, B. *et al.* The impact of incorrect social information on collective wisdom in human groups. *J. R. Soc. Interface* **17**, 20200496 (2020).

14. Mavrodiev, P. & Schweitzer, F. Enhanced or distorted wisdom of crowds? An agent-based model of opinion formation under social influence. *Swarm Intell.* **15**, 31–46 (2021).

15. Poli, R., Valeriani, D. & Cinel, C. Collaborative Brain-Computer Interface for Aiding Decision-Making. *PLoS ONE* **9**, e102693 (2014).

16. Bhattacharyya, S., Valeriani, D., Cinel, C., Citi, L. & Poli, R. 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)* 3099–3102 (IEEE, Berlin, Germany, 2019). doi:10.1109/EMBC.2019.8856309.

17. Ma, S. *et al.* "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* 1–20 (Association for Computing Machinery, New York, NY, USA, 2024). doi:10.1145/3613904.3642671.

18. Arshad, S. Z., Zhou, J., Bridon, C., Chen, F. & Wang, Y. 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 352–360 (Association for Computing Machinery, New York, NY, USA, 2015). doi:10.1145/2838739.2838753.

19. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8**, (2014).

20. Wokke, M. E., Cleeremans, A. & Ridderinkhof, K. R. Sure I'm Sure: Prefrontal Oscillations Support Metacognitive Monitoring of Decision Making. *J. Neurosci.* **37**, 781–789 (2017).

21. Wokke, M. E., Achoui, D. & Cleeremans, A. Action information contributes to metacognitive decision-making. *Sci. Rep.* **10**, 3632 (2020).

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

23. Di Gregorio, F., Maier, M. E. & Steinhauser, M. Are errors detected before they occur? Early error sensations revealed by metacognitive judgments on the timing of error awareness. *Conscious. Cogn.* **77**, 102857 (2020).

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

25. Schnuerch, R., Schmuck, J. & Gibbons, H. Cortical oscillations and event-related brain potentials during the preparation and execution of deceptive behavior. *Psychophysiology* **61**, e14695 (2024).

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

27. Eisma, J., Rawls, E., Long, S., Mach, R. & Lamm, C. Frontal midline theta differentiates separate cognitive control strategies while still generalizing the need for cognitive control. *Sci. Rep.* **11**, 14641 (2021).

28. Cohen, M. X. & Donner, T. H. Midfrontal conflict-related theta-band power reflects neural oscillations that predict behavior. *J. Neurophysiol.* **110**, 2752–2763 (2013).

29. Cohen, M. X. A neural microcircuit for cognitive conflict detection and signaling. *Trends Neurosci.* **37**, 480–490 (2014).

30. Potts, G. F., Martin, L. E., Kamp, S.-M. & Donchin, E. Neural response to action and reward prediction errors: Comparing the error-related negativity to behavioral errors and the feedback-related negativity to reward prediction violations. *Psychophysiology* **48**, 218–228 (2011).

31. Cao, D., Li, Y. & Niznikiewicz, M. A. Neural characteristics of cognitive reappraisal success and failure: An ERP study. *Brain Behav.* **10**, e01584 (2020).

32. McDonnell, A. S. *et al.* This Is Your Brain on Autopilot: Neural Indices of Driver Workload and Engagement During Partial Vehicle Automation. *Hum. Factors* **65**, 1435–1450 (2023).

33. Choe, J., Coffman, B. A., Bergstedt, D. T., Ziegler, M. D. & Phillips, M. E. Transcranial Direct Current Stimulation Modulates Neuronal Activity and Learning in Pilot Training. *Front. Hum. Neurosci.* **10**, (2016).

34. Kilavik, B. E., Zaepffel, M., Brovelli, A., MacKay, W. A. & Riehle, A. The ups and downs of beta oscillations in sensorimotor cortex. *Exp. Neurol.* **245**, 15–26 (2013).

35. Tan, H., Wade, C. & Brown, P. Post-Movement Beta Activity in Sensorimotor Cortex Indexes Confidence in the Estimations from Internal Models. *J. Neurosci.* **36**, 1516–1528 (2016).

36. Parker, A. *et al.* Role of uncertainty in sensorimotor control. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **357**, 1137–1145 (2002).

37. Palmer, C. E., Auksztulewicz, R., Ondobaka, S. & Kilner, J. M. Sensorimotor beta power reflects the precision-weighting afforded to sensory prediction errors. *NeuroImage* **200**, 59–71 (2019).

38. Barratt, E. S. Impulsiveness Defined Within a Systems Model of Personality. in *Advances in Personality Assessment* (Routledge, 1985).

39. Kothe, C. *et al.* The Lab Streaming Layer for Synchronized Multimodal Recording. *bioRxiv* 2024.02.13.580071 (2024) doi:10.1101/2024.02.13.580071.

40. Gramfort, A. *et al.* MEG and EEG data analysis with MNE-python. *Front. Neurosci.* **7**, 267 (2013).

41. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

42. van Rossum, G. & de Boer, J. Interactively testing remote servers using the Python programming language. *CWI Q.* **4**, 283–304 (1991).

43. McKinney, W. Data Structures for Statistical Computing in Python. in 56–61 (Austin, Texas, 2010). doi:10.25080/Majora-92bf1922-00a.

44. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

45. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

46. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).