

Human-Machine Performance Envelope: Controller Adaptive Digital Assistant evaluation

Stefano Bonelli, Anna Vicario, Alfonso Levantesi,
Silvia Torsi
Deep Blue
Rome, Italy

Brais Iglesias, Juan Besada
Universidad Politécnica de Madrid
Madrid, Spain

Christophe Hurter, Alexandre Veyrié
Fédération ENAC-ISAE SUPAERO-ONERA
Toulouse, France

Raquel García Lasheras, Patricia López De Frutos
CRIDA
Madrid, Spain

Gianluca Borghini
BrainSigns
Rome, Italy

José Cañas
Universidad de Granada
Granada, España

Abstract— Air traffic control operators face sustained cognitive demands that can elevate stress, induce fatigue, and impair vigilance. Integrating assistant tools can help alleviate rising workload levels, enabling air traffic control operators to manage increasing traffic volumes while maintaining operational performance. This work employs the Human-Machine Performance Envelope framework designed to capture and assess the dynamic interaction between air traffic control operators and support artificial intelligence systems. To demonstrate its applicability, the Human-Machine Performance Envelope is applied to the CODA (COntroller adaptive Digital Assistant) system, a prototype digital assistant developed to proactively support the activity of the air traffic controllers and reduce mental workload. CODA operates through three functional stages—real-time data ingestion and short-term prediction, context-aware automation adaptation, and dynamic flight control allocation—facilitating adaptive sharing of responsibilities between human operators and automation. Using the Human-Machine Performance Envelope framework, we evaluate the integrated solution's impact on key cognitive and operational metrics, including stress, fatigue, vigilance, and human-AI coordination. Results indicate that the Human-Machine Performance Envelope reliably captures both objective and subjective aspects of controller-AI interaction.

Keywords- *adaptive automation; mental workload; task workload; human-machine interaction;*

I. INTRODUCTION

Air traffic control (ATC) is widely recognized as one of the most cognitively demanding and high-stress occupations. The continuous growth of global air traffic has resulted in increasingly congested airspace, affecting mental workload experienced by Air Traffic Control Operators (ATCOs), and increasing the risk of human error [1]. Elevated stress, fatigue,

and lapses in vigilance not only affect ATCOs well-being but can also degrade operational efficiency and compromise aviation safety [2]. Therefore, reducing the mental workload of ATCOs through the integration of automation remains a critical research priority in air traffic management (ATM) [3]. This study investigates a novel approach for the assessment of the interaction between air traffic control operators and support artificial intelligence systems, the Human-Machine Performance Envelope (HMPE). To exemplify its use, it is used to assess an AI-driven automation system (CODA) designed to proactively support ATCOs through context-sensitive actions and the execution of selected control tasks. The assistant operates within an adaptive automation (AA) framework, whose effectiveness in mitigating controller workload has been demonstrated in previous research [4]. A prototype was developed to process real-time, context-aware information, enabling more efficient and flexible task sharing between human operators and automated functions. The experimental evaluation focused on three main objectives: assessing the validity of the HMPE as a quantitative framework for measuring the variables related to human-AI teaming, examining the quality of human-machine interaction through its indicators, and evaluating the overall impact of the CODA system on ATCOs performance. By analyzing these dimensions, the study assesses the feasibility and potential benefits of the proposed performance evaluation framework and of the CODA system, and contributes shedding light on the impact on adaptive automation systems actual impact in ATM.

The paper starts with an overview of the current state of the art in automation and human-AI teaming within ATC. It then introduces the CODA architecture, followed by the



experimental setup, findings and discussion. Finally, it concludes by outlining the implications of the results and future research directions.

II. RELATED WORK

The integration of artificial intelligence into ATC has prompted a growing body of research aimed at understanding how automation can support human operators in safety-critical environments. One of the earliest and most promising directions has been that of neuroergonomics, which focuses on monitoring cognitive states to inform adaptive automation strategies. Aricò et al. [4] showed that EEG-based passive brain-computer interfaces can be involved in the dynamic activation of automation as a response to increasing mental workload, supporting sustained operator performance in realistic ATC scenarios. Building on these foundations, more recent work has explored the potential for autonomous decision-making agents in ATC. Brittain and Wei [5] introduced a deep multi-agent reinforcement learning approach capable of resolving en-route aircrafts conflicts with high reliability in simulation environments. Their results indicated that autonomy in ATC, when trained on large-scale data and appropriately constrained, can match or even surpass human performance in specific tactical domains. Further advancing this line of inquiry, Vouros et al [6] developed multi-agent graph convolutional models that incorporate domain-specific features to improve transparency, robustness, and explainability in automated conflict resolution. In parallel, Wang et al. [7] emphasized the importance of developing reinforcement learning methods that are both explainable and safety-aware, proposing architectures for autonomous air mobility that embed safety constraints directly into policy learning to enhance transparency and human-machine trust. Yet, the introduction of AI into ATC profoundly restructures human cognition and decision authority. As such, growing literature has focused on the human-automation teaming perspective. Besides, the need for achieving calibrated trust [8][9], increases the choice of adaptive interfaces, grounded in real-time feedback mechanisms and explainability [10]. Endsley [11] underscores that autonomy should not erode operators' situational awareness but rather sustain it through thoughtful information management and interaction design. Despite these advances, a fully integrated approach, one that combines neurophysiological monitoring, adaptive automation, human-AI teaming, and explainable autonomy, remains underdeveloped. Most existing studies focus on isolated dimensions, such as workload adaptation or conflict resolution, without capturing the complex interplay between cognitive states, trust dynamics, and automation behavior in realistic settings. Moreover, few attempts have been made to formalize a bounded cognitive space, such as the Human-Machine Performance Envelope (HMPE) that was adopted in the current study, to guide adaptive strategies. The present work addresses this gap by uniting neurophysiological monitoring with an explainable AI-supported interface, aiming to assess how systems like CODA can maintain ATCOs within safe and efficient cognitive limits. Through EEG, EDA,

and PPG data, alongside performance and subjective metrics, we explore the impact of AI-enabled support on workload, stress, attention, and trust. With this work we contribute to a broader understanding of how AI can support human cognition in the future of air traffic management.

III. CODA SYSTEM DESCRIPTION

CODA is a closed-loop adaptive system that monitors operational and cognitive parameters to support air traffic controllers in maintaining optimal performance. The system integrates real-time data from neurophysiological, behavioral, and operational sources (Subsections A, B, and C) to dynamically adjust automation levels (Subsection D) and allocate part of the incoming workload for ATCO (Subsection E). Figure 1 depicts a high-level architecture of the system. The following subsections describe deeper each of the modules that compose the CODA system.

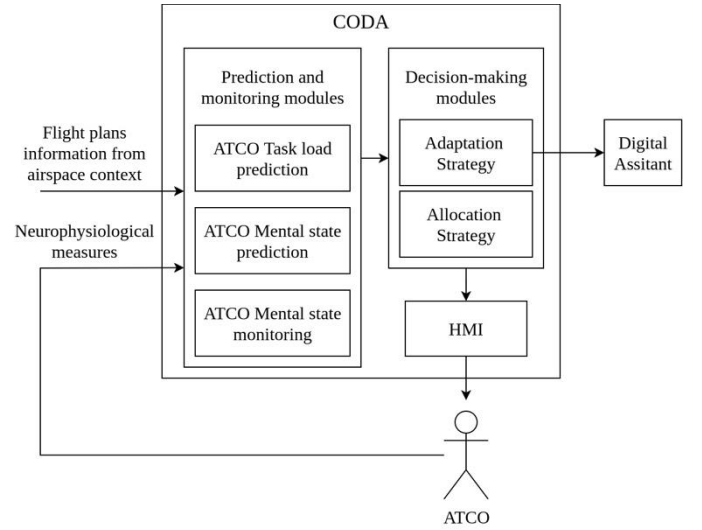


Figure 1. High-level CODA architecture.

A. ATCO Mental states monitoring

Measuring mental states is critical to understanding human cognitive performance, especially in human performance related to demanding environments such as control rooms. In recent years, neurophysiological technologies have enabled significant advances by providing accurate and non-invasive methods to monitor these states through electric signals, autonomic nervous system responses, and analysis of eye behaviors. Among the technologies and sensors explored and tested in CODA, the *Mindtooth Touch* for Electroencephalogram (EEG), and the *Shimmer3 GSR+* for Photoplethysmography (PPG) and Skin Conductance (SC) data collection have been selected. Figure 2 depicts both selected devices during a simulation exercise. The mental states for controllers' performance assessment were: **Mental Workload**, **Stress**, **Vigilance** (Sustained Attention), and **Mental Fatigue**. These indicators were provided in real-time during the experiments as key input for the system decision-making and performance assessment.



Figure 2. Hardware set up for simulations exercise.

B. ATCO Mental states prediction

Apart from the current mental state of the ATCO, future information about its mental state is considered to have key insights for adaptive automation decisions. A cognitive prediction model is designed to forecast the mental states of ATCOs to enhance human-machine collaboration in air traffic management. The model integrates psychological and neuroscientific foundations [12][13] to predict the same four key mental states being monitored. It is structured around two core components: the cognitive structure of ATCOs and the mental resources required to perform tasks. Cognitive structure encompasses processes such as perception, comprehension, projection, and decision-making. Mental resources are categorized into demanded and available resources, forming the basis for evaluating mental state balance.

A central feature of the model is the Cognitive Complexity Tool, which quantifies task difficulty based on traffic patterns and controller mental abstractions (e.g., standard flows, critical points, and groupings). This tool informs the estimation of demanded resources. The model also incorporates Operational Modes (OMs) strategic behavioral patterns that ATCOs adopt under varying workload conditions [12][14]. These modes range from opportunistic (low traffic) to system shock (extreme traffic), and transitions between them are governed by occupancy thresholds derived from historical data.

The prediction of mental states relies on dynamic algorithms that simulate the evolution of available resources over time, accounting for factors such as time-on-task, task complexity and stressor severity. Fatigue is modeled as a depletion of resources, while stress temporarily increases resource availability through neurophysiological activation. Vigilance decrement is inferred from fatigue levels and task monotony.

C. ATCO task load prediction

To predict short-term ATCO **task-load**, forecasting the near future control events (with associated tasks) is proposed. Given the difficulty of estimating atomic tasks in the dynamic and stressed ATCO environment, a probabilistic estimation per task and aircraft was designed. This predicted probability corresponds to the task occurrence probability within the next minutes from the prediction instant on, $[t, t+\Delta t]$, using a data driven approach exploiting historical data in the interest

sector. It is based on simple and explainable features, such as the time to intersect the sector, type of operation or the expected level changes in a flight plan, and a task taxonomy (with 12 different kinds of tasks/events related to ATCO activities such as contacts/releases, conflict resolutions or procedural actions, etc.). From the probability estimation for each task per flight, with the objective of providing a unique final measure of the global upcoming task workload, these individual estimations are aggregated in a global *taskload* (1). For every task in $\{a_1, \dots, a_{12}\}$ and every flight in $\{f_1, \dots, f_n\}$.

$$taskload = \sum_{i=1}^{12} \sum_{j=1}^n P \left((f_j \text{ demands } a_i)_{[t, t+Dt]} \right) \quad (1)$$

D. Adaptation Strategy

Adaptation strategies have been used historically to support improved assistance for ATCOs. Traditional static automation, while reducing workload, introduces well-known drawbacks such as the *out-of-the-loop* performance problem, degradation of operator skills, reduced situation awareness, and an erosion of trust when automation fails [15][16]. To mitigate these risks, CODA proposed a strategy in which automation levels vary adaptively as a function of both operational demands and cognitive states, thereby preserving human agency while enhancing system resilience [17][18].

This Adaptive Automation Strategy (AAS) relies on three core indexes, obtained from the previous mentioned modules: Traffic complexity **T**; Task-load **D**; Controller mental states **M**. Each index is defined in two temporal dimensions: a current value (**Tc**, **Dc**, **Mc**) and a predicted value (**Tp**, **Dp**, **Mp**). Each variable is discretized into three categorical levels: **low**, **medium**, **high**. The resulting decision space is thus:

$$\Omega = \{Tc, Tp, Dc, Dp, Mc, Mp\}, |\Omega| = 3^6 = 729 \quad (3)$$

This configuration space captures the multidimensional dynamics of traffic, tasks, and human states, offering a rich basis for adaptive allocation within the Human Performance Envelope [19]. To map each state vector ω in Ω to an automation action, we implemented a decision tree classifier. Decision trees were chosen for their interpretability and ability to handle categorical variables. Each leaf node corresponds to an automation class:

- **A0**: No automation
- **A1**: Automation light (partial task delegation)
- **A2**: Automation strong (full task delegation)

A simplified pseudocode representation of the allocation process is given below:

```

if Traffic_current == high and Task_current == high:
    if Mental_predicted == high:
        action = A2 # escalate support
    else:
        action = A1 # partial delegation
elif Traffic_predicted == low and Task_predicted == low:
    action = A0 # maintain manual control
else:
    action = A1 # balanced delegation

```

This logic extends across the full decision tree covering all 729 states. Importantly, each decision was accompanied by an explanatory rule set, such as "High traffic and high task load combined with predicted mental overload $\rightarrow$ strong automation". These explanations were communicated via the human-machine interface, ensuring controllers remain aware of why a given allocation was made, embedding transparency and ergonomic design.

E. Flight Allocation strategy

Once the Adaptation Strategy has decided that automation should be triggered, the flight allocation module determines which of the incoming traffic can be safely assigned to the Digital Assistant (DA), which in our design will ideally take care of those flights without expected operational or safety-related interventions from the ATCO. To accomplish this, an explainable logic of the following three factors is proposed:

- **Flight level indicator (PC1):** a flight level change/occupancy indicator is calculated to measure how likely an aircraft flight level will be changed or how congested the flight level is currently.
- **Standard flow indicator (PC2):** a standard flow-based indicator is calculated to measure if the aircraft is following a standard flow and its congestion, determining how likely an aircraft is to be intervened in its current airway.
- **Minimum distance (PC3):** inverse of the minimum predicted distance to nearby aircraft in the following minutes from the decision-making time.

With these three indicators, a criticality measure (2) is determined so the greater the value is the less safe it would be to assign such aircraft to the DA.

$$criticality = k1 \cdot PC1 + k2 \cdot PC2 + k3 \cdot PC3 \quad (2)$$

This value is computed for incoming aircraft, which are sorted in ascending order, so the least critical (and therefore safer to automate) are assigned to the DA. A maximum criticality threshold ensures complex flights demanding accurate and timely control interventions are not assigned.

F. HMI Design and description.

The CODA Human-Machine Interface (HMI) was developed through a user-centered and iterative design process, ensuring that the introduction of adaptive automation would not have compromised the controller's performance. The design process began with an extensive review of existing digital assistant concepts in ATC and related domains, highlighting recurring challenges such as the risk of operators becoming *out of the loop*. Based on this analysis, a set of use case scenarios were defined, in which supportive automation offered operational benefits without high-criticality or strategic decisions. ATCOs were involved in the design of the HMI to understand how automation could be introduced within the controller's workflow without altering mental representations of traffic or established control strategies.

Then, low-fidelity prototypes were produced and evaluated with additional ATCOs to verify that the interface effectively communicated three essential aspects: (i) which flights were under direct controller responsibility, (ii) which flights were proposed for delegation, and (iii) which were currently managed by the DA. Later, higher-fidelity prototypes were integrated with CODA's prediction, adaptation, and allocation modules. The resulting interface (Figure 3) employs a visualization based on columns that spatially organize the flights currently under controller responsibility, those suggested for delegation, and those managed by the DA. An additional visual element in the HMI is timeline-based representation of anticipated automation activity. On the top right of the screen, mental state indicators derived from EEG, EDA, and PPG signals are displayed in a simplified and non-intrusive format, showing only aggregated trends (e.g., workload, stress, vigilance, and fatigue). This minimalist visualization strategy was chosen to inform decision-making without distracting from operational priorities or revealing unnecessary physiological detail. Along the flow of traffic visualization, explanatory notifications alert on the system decisions are available, presenting concise, human-readable rationales (e.g., "High traffic and predicted overload — delegation suggested") for ensuring system explainability and fostering shared situational awareness.



Figure 3. Final prototype of the CODA HMI.

IV. EXPERIMENTAL APPROACH

In this section, we will describe the evaluation of the developed prototype and the performance evaluation framework, which is the focus of the paper, introducing its main contribution, the HMPE.

A. Metrics and validation approach

The core conceptual framework underpinning the CODA system and specially its evaluation was the HMPE. Being the result of neurophysiological, behavioral, and subjective indicators, it is addressed to monitor the joint performance of the human-AI team. The HMPE provides a unified representation of operator-system interaction dynamics by combining multimodal data into a single, interpretable metric. Specifically, the HMPE index was computed by normalizing measures from the three domains—physiological signals, behavioral performance logs, and self-reports related to human-AI interaction—onto a 0-to-1 scale. Established

impairment thresholds were used to define critical boundaries: workload and stress values below 0.1 or above 0.9, and fatigue values above 0.5, were considered indicative of performance degradation. The HMPE was visualized as a triangular area, whose vertices correspond to the three input domains, with a normalized maximum of 1.299 representing the optimal joint performance between the human and the automation. These integrated measures enabled both individual and comparative evaluations, allowing the CODA system to identify when the human–AI team approached or crossed the safe operational boundaries and to trigger adaptive strategies accordingly. The area of the triangle was calculated by using the expression reported in Figure 4, illustrating the HMPE configuration. After normalization, the closer the area is to 1, the higher the controller’s HMPE score.

The validation followed a multimodal, iterative approach designed to capture both quantitative and qualitative aspects of human–AI teaming. Four professional air traffic controllers (ATCOs) from International Federation of Air Traffic Controllers’ Associations, IFATCA participated in the study. Their operational experience ranged from eight to twenty years, and none had previously used CODA. This condition was deliberately maintained to examine first-time usability, transparency, and spontaneous trust calibration. The number of participants was determined according to well-established usability and validation of research principles. As demonstrated by [23], small expert user groups are sufficient to uncover most of the usability and interaction issues in complex human–machine systems, due to the high information yield per user and the diminishing returns observed beyond that range. Prior to the experiment, all participants received a standardized briefing describing the objectives of the evaluation, the simulation environment, and the neurophysiological instrumentation, without being informed about the internal functioning of the adaptive algorithms to avoid cognitive bias. Each ATCO completed three runs corresponding to different traffic scenarios, and the entire experimental session lasted approximately three hours per participant. Between runs, short breaks were introduced to minimize fatigue and avoid carryover effects. Data collection encompassed EEG, SC, and PPG signals, behavioral logs automatically extracted from the simulation platform, and post-run subjective questionnaires evaluating trust, workload, situation awareness, usability, and teaming quality. All data streams were synchronized via a unified timestamp protocol to ensure precise temporal alignment among cognitive, behavioral, and operational events.

B. Experimental prototype

The experimental prototype was implemented to evaluate the CODA system’s cognitive impact on ATCOs and guide the refinement of its adaptive strategies. The study was conducted at CRIDA facilities, making use of the EUROCONTROL ESCAPE Light Simulation Platform.

Each simulation setup was composed of an operational controller working position with a radar display, two pseudo-pilot stations enabling realistic voice communication, a

supervisory console, and an external monitor presenting the CODA explainable AI (XAI) prototype.

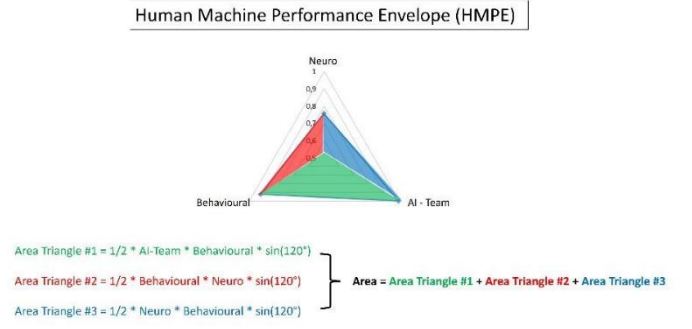


Figure 4. The maximum possible area, corresponding to all indices equal to 1, was 1.299. Individual HMPE scores were therefore normalized by this value.

Before each run, the ATCO underwent a brief familiarization phase and a three-minute baseline recording to measure their physiological measures during rest. During the simulation, the CODA interface provided real-time visual feedback on automation activity, adaptive recommendations, and flight delegation proposals. Each run lasted approximately one hour and twenty minutes and followed a standardized sequence: baseline, warm-up, full-scenario execution, and post-run debriefing. Throughout execution, CODA’s adaptive automation remained active, adjusting its support according to current and predicted task load, traffic complexity, and cognitive state. After each run, participants completed a post-run questionnaire and took part in a semi-structured interview to provide qualitative insights into CODA’s behavior, transparency, and usability. The combination of objective measures and subjective reports enabled triangulation across data modalities, strengthening the interpretation of results. Although the system was not yet fully integrated into the simulator display, its presentation through an external display ensured clear separation between the traditional control interface and the assistant’s adaptive functions, which supported transparency and avoided perceptual interference.

C. Test scenarios description

The experimental scenarios were designed to test the CODA system under varying traffic demands and operational conditions. The simulations took place within Spanish airspace, specifically in the Madrid Area Control Center (ACC), focusing on the Domingo Upper (DGU) sector, shown in Figure 5. The DGU sector manages *en-route* traffic between flight levels FL345 and FL660, with a standard horizontal separation of five nautical miles. This sector was selected for its medium complexity, which allowed controlled variations in task load while maintaining operational realism. The DGU sector serves as a primary transit corridor to and from Adolfo Suárez Madrid–Barajas Airport (LEMD), Spain’s main international hub, and is characterized by mixed flows of northbound and southbound traffic intersecting at multiple waypoints.



Figure 5. Domingo Upper.

Traffic density varied from low to very high, depending on the experimental condition, and three distinct traffic patterns were used to assess CODA's performance across different levels of workload dynamics. The Steady Progression (SP) scenario featured a gradual increase in traffic, allowing assessment of CODA's capability to support workload buildup. The Baseline Perturbation (BP) scenario simulated a progressive reduction in traffic, enabling analysis of recovery and adaptation processes after high-demand phases. The Sudden Burst Application (SBA) scenario presented two pronounced traffic peaks separated by medium-load intervals, testing the system's responsiveness under fluctuating complexity.

V. EXPERIMENTAL RESULTS

This section presents the results of the experimental evaluation of the CODA system, focusing on neurophysiological indicators, behavioral performance, subjective assessments, and the integrative HMPE metrics. Findings are reported at both the individual and scenario levels to capture inter-individual variability and context-dependent effects.

A. Neuro-Cognitive and Physiological Indicators

The neurophysiological data revealed consistent trends across varying cognitive load conditions. In agreement with prior literature [20], increases in frontal-midline theta (4–7 Hz) activity were associated with higher task demands, reflecting enhanced mental effort during complex operational phases. Concurrent elevations in frontal-temporal beta power and electrodermal activity (EDA) were observed during moments of acute stress, suggesting that the system effectively captured arousal-related fluctuations in real time. Indicators of fatigue, including rising delta/theta ratios, reductions in heart rate variability (HRV), and decreases in tonic skin conductance levels, appeared progressively throughout the simulation, reflecting gradual declines in vigilance and cognitive resources. These temporal patterns corresponded closely with event-based logs from the simulator, allowing precise correlation between cognitive dynamics, operational load, and scenario-specific challenges.

B. Human–Machine Performance Envelope (HMPE)

Applying the HMPE framework to each controller operating under CODA-enabled scenarios yielded to the corresponding HMPE values. Figures 6 (a)–(d) illustrate the HMPE triangle for each participant, and Table I summarizes their overall scores.

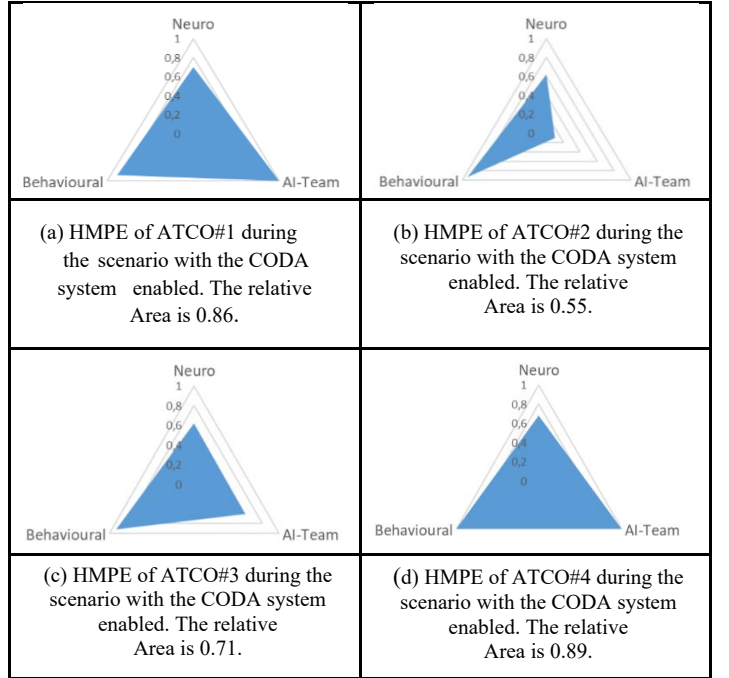


Figure 6. Summary of HMPE score.

ATCOs #1 and #4 achieved higher HMPE scores, reflecting a relatively balanced alignment between cognitive stability, operational performance, and subjective teaming. ATCO #3 obtained an intermediate score (0.71), characterized by greater variability across behavioral and subjective indicators. ATCO #2 recorded the lowest HMPE (0.55), with lower teaming and greater performance dispersion. Overall, CODA contributed to maintaining good performance levels (all values > 0.88), adequate teaming (all > 0.6), and stable mental states (all > 0.62). The lower teaming value for ATCO #2 likely reflects limited familiarization with CODA, as this participant completed only one scenario before data collection was interrupted due to weather-related disruptions during the validation period.

TABLE I. SUMMARY OF THE HMPE SCORE TEAMING SCORES

Controller	ATCO#1	ATCO#2	ATCO#3	ATCO#4
HMPE Score	0.86	0.55	0.71	0.89

C. Post-Run Questionnaire Results

Subjective perceptions of human–AI interaction were collected through a 17-item Post-Run Questionnaire addressing Trust, Transparency, Situation Awareness, Task Allocation, and Usability (Figure 7). Participants rated each item on a 10-point Likert scale, and Table II presents the aggregated results.

TABLE II. AVERAGE TEAMING SCORES BY ATCO.

Controller	ATCO#1	ATCO#2	ATCO#3	ATCO#4
Avg. Teaming Score	6.9 (range: 5.7–8.1)	8.0 (single scenario)	5.6 (range: 5.2–6.0)	6.5 (range: 5.0–7.7)

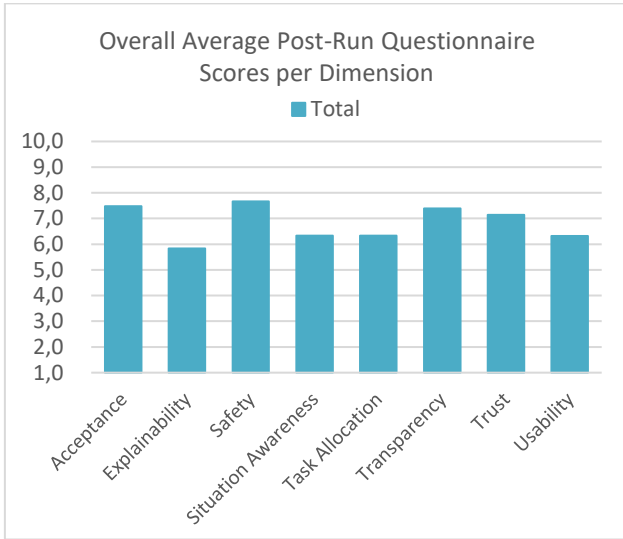


Figure 7. Overall average score for each dimension of the Post-Run Questionnaire, aggregated across all participants and scenarios.

As shown in Figure 8, ATCO #2 reported the highest overall teaming perception (score = 8.0), indicating a strong sense of coordination and trust. ATCO #3 reported the lowest average score (5.6), reflecting some initial adaptation difficulties, consistent with the HMPE outcomes. ATCOs #1 and #4 exhibited more variable patterns across scenarios.

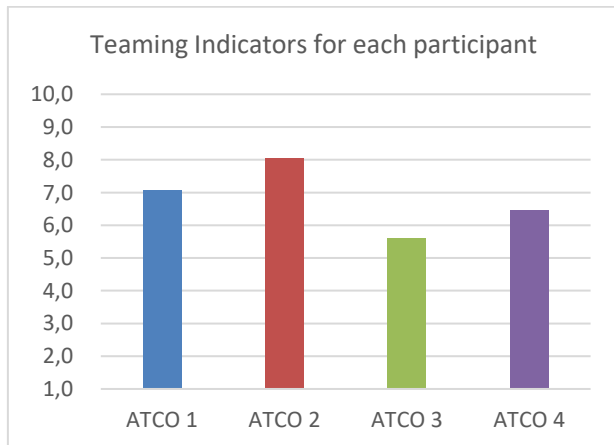


Figure 8. Average Teaming Indicator for each participant.

D. Scenario-Specific Analysis

Scenario type appeared to influence both objective and subjective outcomes. The Baseline Perturbation (BP) scenarios—characterized by progressive decreases in traffic—

elicited the highest teaming ratings, suggesting that moderate workload conditions favored effective human–AI coordination. The Steady Progression (SP) scenarios, involving gradual workload increases, produced greater variability among participants, indicating differences in adaptation strategies and stress responses [21]. The Sudden Burst Application (SBA) scenarios yielded intermediate results, with heterogeneous effects depending on the controller.

TABLE III. TEAMING SCORES BY SCENARIO TYPE.

Scenario	BP	SP	SBA
Avg. Teaming Indicator	Highest overall	Highest variability	Moderate, heterogeneous

The analysis of the Post-Run Questionnaire data revealed significant variations in the Teaming indicator among controllers across different scenarios (Figure 9). ATCO #1 displayed progressive improvement across runs, increasing from 5.7 (SP) to 8.1 (SBA), suggesting a possible learning or adaptation effect. Conversely, ATCO #3 showed a decline (from 6.0 in BP to 5.2 in SBA), consistent with signs of cognitive fatigue or reduced satisfaction with system behavior.

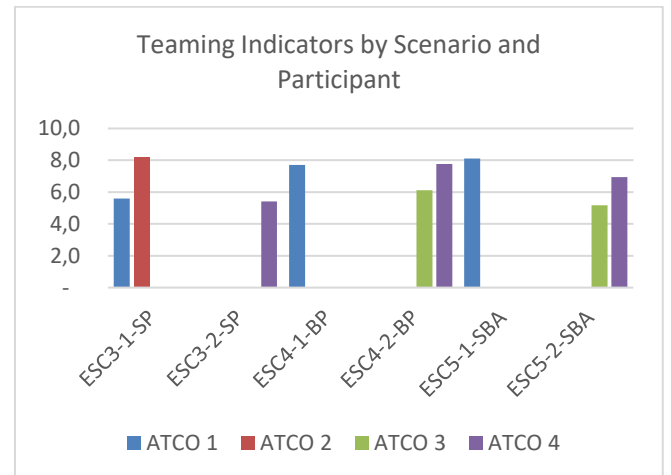


Figure 9. Teaming Indicators trend across different scenarios for each ATCO.

Aggregated results (Figure 10) indicate that BP scenarios were generally rated most favorably, while SP runs revealed the widest individual differences. SBA scenarios were rated moderately, though with a downward trend for ATCO #3. The findings indicate that the perceived effectiveness and quality of collaboration with the CODA system were not uniform but were modulated by both scenario characteristics and inter-individual differences among participants. For example, ATCO #2—who completed only a single run—reported notably high teaming scores, a result that may reflect the specific suitability of the scenario rather than an overall system effect. This variability underscores the crucial

influence of contextual and experiential factors in shaping the outcomes of human–AI interaction. Consistent with recent research, the effectiveness of human–AI teaming depends on the system’s capacity to adapt to operator strategies and dynamic task demands, thereby fostering appropriate trust, transparency, and shared situation awareness [11][24][25].

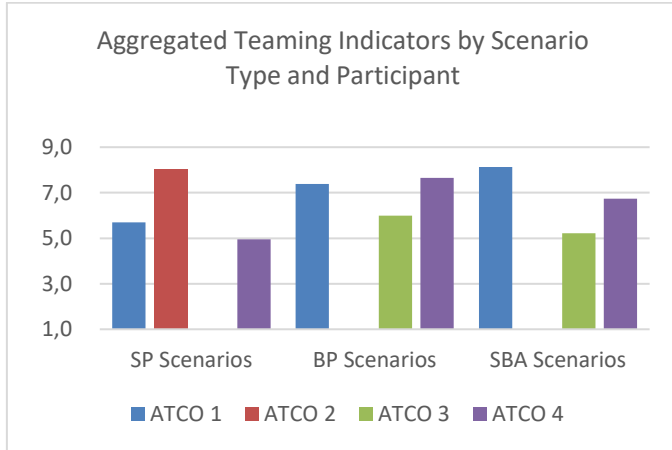


Figure 10. Aggregated Teaming Indicator scores for each scenario type (SP, BP, SBA) and participants.

E. Results Discussion

This study implemented and evaluated the application of the HMPE in the context of task allocation performed by an AI delegation system. Its composite metrics encompass neurophysiological, behavioral, and subjective indicators to quantify human–AI teaming. The normalized data streams were combined into a triangular visualization index, enabling the objective assessment of CODA’s capacity to sustain safe and effective performance across varying operational conditions. HMPE values ranged between 0.55 and 0.89. Higher scores were associated with more stable cognitive states, stronger teaming, and greater system familiarity, whereas lower values corresponded to initial exposure and higher scenario complexity. In fact, subjective reports were consistent with these values: controllers with higher HMPEs also expressed stronger trust, transparency, and collaborative alignment. Differences across scenarios show that the effectiveness of CODA’s adaptive support varies with both the level of traffic complexity and the controller’s familiarity with the system, underscoring the need for context-sensitive adaptation.

Three overarching insights emerged. First, trust calibration appeared essential: CODA’s adaptive transparency promoted appropriate reliance, though variability during early use indicated the need for structured familiarization. Second, situation awareness and cognitive resilience were largely preserved, as the system adapted to workload changes and provided intelligible support, though its effectiveness varied across contexts. Third, explainability proved critical: well-timed explanations enhanced shared understanding, while excessive or poorly timed feedback occasionally brought distraction.

Overall, the HMPE effectively captured both objective and subjective dimensions of controller–AI interaction, supporting its potential as a practical framework for evaluating adaptive automation. The CODA system showed promise in sustaining performance, cognitive stability, and acceptance when appropriately tuned to user needs and context.

VI. CONCLUSIONS AND FUTURE WORK

This study presented and evaluated the HMPE [22] as a framework for assessing cognitive and physiological states in safety-critical domains. Using high-fidelity ATC simulations on the ESCAPE Light platform, four professional ATCOs interacted with the CODA system while their mental states were continuously monitored through neurophysiological sensors (EEG, PPG, EDA). These data, combined with behavioral and subjective metrics, were integrated into a composite HMPE index. Findings suggest that CODA can contribute to sustaining controller performance and cognitive stability, with HMPE scores ranging between 0.55 and 0.89. Higher values were associated with more stable mental states, stronger human–AI teaming, and greater system familiarity, whereas lower values reflected limited exposure and scenario complexity. The results identified three key factors influencing human–AI teaming effectiveness: accurate trust calibration, context-aware adaptive behavior, and efficient explainability. Beyond these empirical results, the study highlights the potential of the HMPE as a diagnostic and design tool for human–AI joint systems. By providing a quantitative view of cognitive states and team dynamics, the HMPE offers a means to monitor and adapt human–machine interaction in real time. This could inspire AI based, adaptive interface management approaches in ATC, enabling automation to dynamically adjust its level of assistance according to operators’ personal and contextual-driven cognitive boundaries rather than fixed operational thresholds. While the results are encouraging, they should be interpreted with caution given the preliminary nature of the adaptive algorithms and the limited sample size, which constrain statistical robustness and generalizability. Future research should involve a broader and more diverse group of controllers and adopt longitudinal designs to examine how trust, workload management, and skill adaptation evolve over sustained interaction with adaptive systems. Embedding the HMPE framework into iterative system development could help identify early signs of cognitive or teaming degradation and inform real-time adjustments to automation strategies. In the longer term, continuous HMPE feedback may enable the design of closed-loop adaptive systems capable of maintaining operators within optimal performance ranges under dynamic operational conditions. At a broader level, future research could adopt a socio-technical perspective by extending the HMPE toward a Team Performance Envelope (TPE) that captures shared awareness, coordination efficiency, communication load, and inter-agent trust among ATCOs,

pilots, and AI systems. Investigating these dynamics in high-fidelity, AI-integrated simulation environments could advance the understanding of how hybrid cognitive ecologies, comprising human operators and intelligent agents, coordinate and sustain resilient performance, providing an empirical basis for the design of transparent, adaptive, and context-aware automation in future air traffic management systems.

ACKNOWLEDGMENT

This work is part of the SESAR CODA project; project type: exploratory research; SESAR program: Digital European Sky; grant ID: 101114765.

REFERENCES

- [1] Hah, S., Willems, B., & Phillips, R. (2006, October). The effect of air traffic increase on controller workload. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 1, pp. 50-54). Sage CA: Los Angeles, CA: SAGE Publications.
- [2] Edwards, T., Gabets, C., Mercer, J., & Bienert, N. (2016, July). Task demand variation in air traffic control: Implications for workload, fatigue, and performance. In *Advances in Human Aspects of Transportation: Proceedings of the AHFE 2016 International Conference on Human Factors in Transportation, July 27-31, 2016, Walt Disney World®, Florida, USA* (pp. 91-102). Cham: Springer International Publishing.
- [3] Pasaoglu, C., Baspinar, B., Ure, N. K., & Inalhan, G. (2016). Hybrid systems modeling and automated air traffic control for three-dimensional separation assurance. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 230(9), 1788-1809.
- [4] Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., ... & Babiloni, F. (2016). *Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment*. *Frontiers in human neuroscience*, 10, 539.
- [5] Brittain, M., & Wei, P. (2019). Autonomous Air Traffic Controller: A Deep MultiAgent Reinforcement Learning Approach. *arXiv preprint arXiv:1912.09143*. <https://arxiv.org/abs/1912.09143>
- [6] Vouras, G. A., Papadopoulos, G., & Salampasis, M. (2022). Automating the Resolution of Flight Conflicts Using Deep Reinforcement Learning and DomainKnowledgeEnhanced Graph Networks. *arXiv preprint arXiv:2211.06614*. <https://arxiv.org/abs/2211.06614>
- [7] Wang, L., Yang, H., Shao, K., & Liu, Y. (2022). Explainable and Safe Reinforcement Learning for Autonomous Air Mobility. *arXiv preprint arXiv:2212.03069*. <https://arxiv.org/abs/2212.03069>
- [8] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80. <https://doi.org/10.1518/hfes.46.1.50.30392>
- [9] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), 697-718. [https://doi.org/10.1016/S10715819\(03\)000387](https://doi.org/10.1016/S10715819(03)000387)
- [10] Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407-434. <https://doi.org/10.1177/0018720814547570>
- [11] Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors*, 59(1), 5-27. <https://doi.org/10.1177/0018720816681350>
- [12] Histon, J. M., & Hansman, R. J. (2008). *Mitigating complexity in air traffic control: the role of structure-based abstractions* (Doctoral dissertation, Massachusetts Institute of Technology, Department of Aeronautics and Astronautics).
- [13] Hockey, G.R.J., Coles, M.G.H., Gaillard, A.W.K. (1986). *Energetical Issues in Research on Human Information Processing*. In: Hockey, G.R.J., Gaillard, A.W.K., Coles, M.G.H. (eds) *Energetics and Human Information Processing*. NATO ASI Series, vol 31. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-4448-0_1
- [14] Cho, A; Histon, J.M.; de Albuquerque Filho, E.a.f.; Hasman, R.J. (2011). *Identification of Air Traffic Control Sectors with Common Structural Features*. 2011 16th International Symposium on Aviation.
- [15] M. R. Endsley and E. O. Kiris, "The out-of-the-loop performance problem and level of control in automation," *Human Factors*, vol. 37, no. 2, pp. 381-394, 1995. doi: 10.1518/001872095779064555.
- [16] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230-253, 1997. doi: 10.1518/001872097778543886.
- [17] D. B. Kaber and M. R. Endsley, "The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task," *Theoretical Issues in Ergonomics Science*, vol. 5, no. 2, pp. 113-153, 2004. doi: 10.1080/1463922021000054335.
- [18] D. Bernabei and F. Costantino, "Adaptive automation: Definitions, frameworks, and applications in human-machine interaction," *Cognitive Systems Research*, vol. 74, pp. 14-28, 2022. doi: 10.1016/j.cogsys.2022.02.004.
- [19] P. Aricò, G. Borghini, G. Di Flumeri, N. Sciaraffa, and F. Babiloni, "Passive BCI beyond the lab: Current trends and future directions," *Physiology*, vol. 31, no. 4, pp. 358-373, 2016. doi: 10.1152/physiol.00004.2016.
- [20] Fairclough, S. H., Moores, L. J., Ewing, K. C., & Roberts, J. (2009, September). Measuring task engagement as an input to physiological computing. In *2009 3rd international conference on affective computing and intelligent interaction and workshops* (pp. 1-9). IEEE.
- [21] Wickens, C. M., Wiesenath, D. L., & Roseborough, J. E. (2015). In situ methodology for studying state driver stress: A between-subjects design replication. *Journal of Applied Biobehavioral Research*, 20(1), 37-51.
- [22] Graziani, I., Berberian, B., Kirwan, B., Le Blaye, P., Napoletano, L., Rognin, L., & Silvagni, S. (2016, September). Development of the human performance envelope concept for cockpit HMI design. In *HCI-Aero 2016 International Conference on Human-Computer Interaction in Aerospace*.
- [23] Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (pp. 206-213). ACM
- [24] Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. A., & Hancock, P. A. (2022). Teamwork, trust, and communication in human-autonomy teaming: A meta-analysis and empirical review. *Human Factors*, 64(3), 417-439.
- [25] Schmutz, J. B., Kolbe, M., Grote, G., & Manser, T. (2024). Human-AI teaming in safety-critical environments: The role of shared situation awareness, trust, and adaptive coordination. *Cognition, Technology & Work*, 26(1), 45-62.