The Aeronautical Journal (2026), **130**, pp. 2808–2830
doi:10.1017/aer.2026.10182

RESEARCH ARTICLE

Sensitive physiological feature analysis related to human–automation trust assessment and discrimination

Chuanyan Feng[1,2,3], Shuang Liu[2], Xiaoru Wanyan[2], Chunying Qian[2], Yuqing Dang[2], Yangrui Lai[2] and Yaoming Zhou[2,3]

[1]Hangzhou International Innovation Institute, Beihang University, Hangzhou, China
[2]School of Aeronautical Science and Engineering, Beihang University, Beijing, China
[3]Tianmushan Laboratory, Hangzhou, China
Corresponding author: Xiaoru Wanyan; Email: wanyanxiaoru@buaa.edu.cn

Received: 13 November 2025 UTC; **Revised:** 23 April 2026 UTC; **Accepted:** 23 April 2026 UTC

Keywords: human–automation trust; EEG; fNIRS; interpretability modeling; trust discrimination

Abstract

This study aims to identify physiological features sensitive to human–automation trust (HAT) and to distinguish between high HAT (HHAT) and low HAT (LHAT) levels. A simulated aviation monitoring and takeover task was conducted under a 2 (reliability) × 2 (transparency) experimental design, in which HHAT and LHAT levels were defined based on their Jian scale scores. Paired *t*-tests were then employed to analyse HAT-sensitive features derived from five physiological modalities: electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), eye tracking, facial expression and electrocardiography (ECG). Finally, HAT discrimination models with varying interpretability were constructed using the identified sensitive features. The results indicated that: (1) compared with HHAT, LHAT exhibited significantly higher EEG α relative power and lower δ relative power, shorter eye-tracking metrics including the average duration of whole fixations and the average and total durations of fixations in takeover and flight areas, as well as higher ECG metrics such as the standard deviation of normal-to-normal intervals, the low-frequency to high-frequency ratio, and the low-frequency ratio; (2) no significant differences were found in fNIRS oxygenated haemoglobin and deoxygenated haemoglobin concentrations or in facial micro-expression and action units across HAT levels and (3) the highly interpretable linear support vector machine model constructed based on HAT-sensitive features achieved optimal performance, yielding a fivefold cross-validation accuracy of 0.8065 and an F1-score of 0.7942. These findings can provide empirical evidence for identifying HAT-sensitive physiological markers and developing real-time HAT assessment methods, thereby supporting the monitoring of trust degradation that may compromise flight safety.

Nomenclature

HAT	Human–Automation Trust
KNN	K-Nearest Neighbors
RF	Random Forest
LDA	Linear Discriminant Analysis
LR	Logistic Regression
EEG	Electroencephalogram
fNIRS	functional Near-infrared Spectroscopy
α	Alpha
β	Beta
γ	Gamma
δ	Delta
θ	Theta
HbO	Oxygenated Hemoglobin

© The Author(s), 2026. Published by Cambridge University Press on behalf of Royal Aeronautical Society.

DLPFC	Dorsolateral Prefrontal Cortex
DMPFC	Dorsal Medial Prefrontal Cortex
HbR	Deoxygenated Hemoglobin
ECG	Electrocardiography
HRV	Heart Rate Variability
LF/HF	Low-Frequency/High-Frequency
HR	Heart Rate
SDNN	Standard Deviation of Normal to Normal
ADAS	Advanced Driver Assistance Systems
GSR	Galvanic Skin Response
CNN-LTSM	Convolutional Neural Networks with Long Short-Term Memory networks
LightGBM	Light Gradient Boosting Machine
XGBoost	Xtreme Gradient Boosting
QDA	Quadratic Discriminant Analysis
SYS	SYStem monitoring
AF-MATB	Air Force Multi-Attribute Task Battery
HHAT	High HAT
LHAT	Low HAT
MATB II	Multi-Attribute Task Battery II
TAK	TAKeover and flight
RES	RESource regulation and pump monitoring
TWS	Takeover Warning System
ICA	Independent Component Analysis
FFT	Fast Fourier Transform
ADWF	Average Duration of Whole Fixations
ADF.TAK	Average Duration of Fixations in the TAK area
TDF.TAK	Total Duration of Fixations in the TAK area
NF.TAK	Number of Fixations in the TAK area
AAS	Average Amplitude of Saccade
APVS	Average Peak Velocity of Saccades
SDPVS	Standard Deviation of Peak Velocity of Saccades
AWFPD	Average Whole-Fixation Pupil Diameter
AU	Action Unit
BPM	Beats Per Minute
LF	Low Frequency
TP	Total Power
LF ratio	Low Frequency ratio
HF ratio	High Frequency ratio
LF_nu	Low Frequency in normalized units
ANOVA	ANalysis Of VAriance
PCA	Principal Component Analysis
DT	Decision Tree
GNB	Gaussian Naive Bayes
MLP	Multilayer Perceptron
SVM	Support Vector Machine
fMRI	functional magnetic resonance imaging

1.0 Introduction

In the field of human–computer interaction, human–automation trust (HAT) is generally defined as 'the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability' [1]. Between 2018 and 2019, two major aviation accidents involving the Boeing 737 MAX, namely Lion Air Flight 610 and Ethiopian Airlines Flight 302, resulted in a total of 346 fatalities, serving as a tragic illustration of the catastrophic consequences arising from a breakdown

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

of HAT. Investigations revealed that the direct cause of these accidents was the manoeuvring characteristics augmentation system, which operated silently in the background with unpredictable behaviour and poor transparency in its design [2, 3]. This design flaw prevented pilots from effectively delegating decision-making and control authority to the automated system in critical situations, ultimately leading to a collapse of trust, escalated human–automation conflict and the loss of both aircraft and lives. As interactions between human operators and intelligent agents become increasingly frequent, the measurement and discrimination of HAT have become ever more crucial, yet related research remains relatively underdeveloped. The quantification of trust using multidimensional measures encompassing subjective, behavioural and physiological dimensions enables the detection of trust biases such as overtrust or distrust, thereby optimising human–machine interface design, informing targeted training strategies and ultimately ensuring safe and efficient human–automation collaboration in complex aviation environments.

Recent studies have identified three primary approaches for measuring HAT: self-report, behavioural and physiological measures [4], each exhibiting distinct strengths and limitations in application [5]. Self-report measures, such as the Jian scale [6], have the advantages of being easy to implement, cost-effective and exhibiting relatively high face validity. However, these approaches typically rely on retrospective ratings conducted after task completion; they are susceptible to recall bias and therefore tend to be highly subjective. Behavioural measures are objective and non-intrusive, yet inferring trust levels solely from behavioural indicators such as task completion time or error rates remains contentious, as trust is not the only factor influencing behaviour [5]. In comparison, physiological measures have attracted growing attention for their real-time responsiveness and objectivity, despite requiring specialised equipment and rigorous experimental control. Regarding physiological features, the sensor suite for the flight environment has gained significant interest. Given the constraints of aviation dynamic scenarios, non-contact approaches show great promise. While the minimal requirements for sensor configuration have not yet been clearly defined, the selection and combination of limited physiological features can result in markedly different trust detection outcomes. Nevertheless, exploring the effects of various physiological feature combinations remains valuable. The exploration of sensitive physiological markers has emerged as one of the fastest-growing areas in HAT research [5].

Furthermore, the discrimination of HAT-related states has attracted considerable attention, primarily employing supervised learning, reinforcement learning and related techniques for estimating or classifying trust levels [7]. Supervised learning methods typically include models with limited interpretability, such as k-nearest neighbours (KNN) and random forest (RF), as well as more interpretable models, such as linear discriminant analysis (LDA) and logistic regression (LR). The discriminative performance varies depending on the interpretability of the chosen model. In contrast, reinforcement learning methods are generally less interpretable, yet they often achieve superior discrimination performance. Currently, the discrimination of HAT-related states in aviation environments not only lacks comprehensive integration of the aforementioned sensitive features but also suffers from insufficient comparative analysis regarding algorithm interpretability, highlighting the need for further investigation.

1.1 Literature review

In the exploration of sensitive indicators related to HAT-related states. Current physiological approaches for capturing HAT fall into two broad categories: neural measures and various types of physiological responses [5]. Within the domain of neural activity assessment, electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) are two representative techniques. EEG provides millisecond-level temporal resolution, allowing precise tracking of rapid fluctuations in cortical activity [8]. A substantial body of evidence suggests that statistically significant neural signatures are often concentrated in rhythmic activities within specific frequency bands [9]. For instance, during decision-making and word elicitation tasks, trusted scenarios are associated with increased alpha (α) and beta (β) power in the frontal lobes, indicating active neural connectivity, whereas distrusted scenarios elicit elevated gamma (γ) activity in the temporal lobes [10, 11]. In highly automated driving scenarios,

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

suppressed frontal α activity has been observed during system malfunctions [12], while increased β power has been linked to lower system trust [8]. Similarly, in scaffolding and rectification tasks, high α power has been associated with greater trust levels [13]. Moreover, findings from investment game paradigms indicate that trust and trust-related decisions are closely associated with delta (δ) and γ activity [9], as well as α suppression at parietal electrode sites [14], suggesting that high-frequency oscillations and inhibitory rhythms may serve as effective neural markers of trust.

fNIRS, a non-invasive brain imaging technique, offers moderate spatial resolution, strong resistance to motion artifacts and high portability, making it ideal for HAT studies in naturalistic or high-ecological validity environments. Empirical evidence suggests that changes in oxygenated haemoglobin (HbO) concentration measured by fNIRS may serve as a sensitive indicator of trust states. For instance, during non-driving-related tasks in highly automated driving contexts, the group with low credibility showed increased HbO concentration in the dorsolateral prefrontal cortex (DLPFC) during risk events [15]. In crime mapping tasks, elevated HbO activity was observed in the dorsal medial prefrontal cortex (DMPFC) prior to low-trust events, whereas deoxygenated haemoglobin (HbR) did not show a significant difference [16]. Similarly, in an image classification task with an AI teammate, higher concentrations of both HbO and HbR were detected under trusted conditions compared to distrusted ones [17]. In collaborative robot scenarios, erroneous robot behaviour elicited enhanced activation in the medial and right dorsolateral prefrontal cortices [18]. These findings suggest that neural responses measured by EEG and fNIRS reflect variations in trust; however, current research remains largely confined to domains such as automated driving and human–robot collaboration, with sensitive features mainly centred on EEG α activity and fNIRS HbO. Investigations into the neural mechanisms underlying HAT in aviation-specific contexts remain limited, highlighting the need for further research to extend and refine these findings.

Various physiological responses are commonly used to assess HAT, including eye tracking, facial expression analysis and electrocardiography (ECG) monitoring. As non-invasive approaches, eye tracking and facial expression analysis are frequently employed to evaluate HAT. Eye tracking offers the advantage of being minimally intrusive and providing real-time insights into visual attention allocation. Nonetheless, studies have yielded inconsistent findings regarding its relationship with HAT, around half of the studies identified significant associations, primarily related to fixation metrics, while others found no clear patterns [19]. For example, in scenarios involving highly automated driving, self-reported trust was associated with reduced monitoring of the automated system; during a UAV-based military target identification task, low-reliability conditions resulted in longer fixation durations, more frequent fixations and extended scan paths [20]; in human–robot collaborative assembly tasks, the trust-breach phase showed increased fixation counts and decreased gaze transition entropy compared to the trust-building phase [21]; and in a planetary gear system cooperation task, unreliable robot behaviour led to increased gaze frequency towards the robot, higher stationary gaze entropy and reduced gaze transition entropy [22].

Facial expression analysis captures micro-expressions associated with trust; however, research linking facial expressions to trust remains in its early stages. For instance, in a competitive card-bluffing game, trusted states were associated with increased facial actions such as ‘happy’, ‘mouth smile left’ and ‘mouth smile right’ [23]. However, facial expressions may be influenced by contextual factors and exhibit significant individual variability, which necessitates further validation of their reliability as standalone indicators of trust. ECG has been shown to reflect trust-related autonomic activation, with elevated heart rate variability (HRV) observed under unreliable conditions [24, 25], though some studies show no consistent patterns [26]. For instance, during a metal-surface polishing task, unreliable behaviour from a UR10 collaborative robot triggered a significant decrease in the low-frequency/high-frequency (LF/HF) ratio, suggesting a suppression of sympathetic activity [25]; conversely, in a dynamic trust-induction task involving planetary gear assembly, no significant changes were observed in heart rate (HR), standard deviation of normal to normal (SDNN), or LF/HF [21], indicating that trust fluctuations do not always evoke explicit ECG responses. In contrast, in a competitive bluff game task, high-trust states were linked to elevated mean HR, potentially reflecting physiological activation during engaged participation

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

[27]. In summary, various HAT-sensitive physiological responses are primarily reflected in gaze-based metrics from eye tracking, micro-expressions from facial analysis and HRV-related indices from ECG. However, these findings require further exploration and validation, particularly within aviation-specific contexts.

In the field of trust discrimination research, researchers have developed various trust models [8], predominantly utilising supervised learning, reinforcement learning and related techniques for estimating or classifying trust levels [7]. For instance, in a real-road experiment, leveraging inputs such as hand position transition probability matrices, age, gender and advanced driver assistance systems (ADAS) types, an RF algorithm achieved 79.57% accuracy in discriminating three trust levels derived from k-means clustering of self-reported scores [28]. In driving simulator takeover tasks, RF was employed to analyse physiological features and takeover-related factors for binary classification of trust-increasing and trust-decreasing groups, achieving 73.7% accuracy and an 77.3% F1-score [29]; a subsequent study using a smoothed convolutional neural networks with long short-term memory networks (CNN-LTSM) method maintained 73.7% accuracy and a 75.3% F1-score [30]. Similarly, another study classified driver trust into low, medium and high levels based on self-reported ratings and by integrating whole-brain features with a light gradient boosting machine (LightGBM) algorithm, achieved an accuracy of 88.44% and an F1-score of 78.31% [8]. In an image classification task, self-reported ratings were used to define trust and mistrust conditions, with a 200-dimensional feature vector constructed from five characteristics (mean, standard deviation, kurtosis, minimum and maximum) of fNIRS HbO and HbR across 20 channels, and an Xtreme Gradient Boosting (XGBoost) model achieved the highest accuracy of 76.0% [17]. In a VR-based collaborative bricklaying task, features including fNIRS, GSR, HR and head motion were used as inputs to a convolutional neural network (CNN) model for classifying trust-decreasing and trust-increasing states, with an accuracy of 71.9% [31]. Another study conducted in a chemical industry scenario, where a collaborative robot (cobot) assisted in handing over a beaker and pouring chemicals, used a 4D tensor (batch size, height, width and depth) as input, and a CNN model classified high and low trust, achieving accuracies of 78.61% in the handing task and 73.35% in the pouring task [32]. Additional studies also explored multimodal inputs, including Galvanic Skin Response (GSR), HR and eye-tracking features; 28×28 pixel time–frequency–spatial images from EEG signals; 149-dimensional feature vectors from EEG and GSR; and eye-tracking fixation/saccade metrics, combined with classifiers such as XGBoost [24], CNN [33], quadratic discriminant analysis (QDA) [34] and RF [35, 36] in diverse scenarios including the SYStem monitor (SYS) sub-task of Air Force Multi-Attribute Task Battery (AF-MATB), driving takeover tasks, unmanned aerial vehicle (UAV) control and human–agent monitoring tasks, with reported accuracies ranging from 71.9% to 92.97%. Notably, while low-interpretability models such as CNN have been applied to trust discrimination, their performance does not necessarily surpass that of traditional machine learning methods with high interpretability, such as LR, and they often require substantial computational resources and large datasets, limiting their real-time applicability [8]. Overall, most existing trust discrimination models exhibit accuracies below 80%, and there remains a lack of solutions that combine high interpretability with precision, particularly in aviation contexts where online trust monitoring based on sensitive features remains a critical research focus.

1.2 Research objective

Based on the background and requirements outlined above, this study aims to explore HAT-sensitive physiological features and construct a model to discriminate HAT levels. A simulated aviation monitoring and takeover task was designed with a 2 (reliability) $\times$ 2 (transparency) factorial structure. Two distinct HAT conditions, High HAT (HHAT) and Low HAT (LHAT), were classified based on their Jian scale scores. Subsequently, the paired t-test method was used to systematically analyse the HAT-sensitive features in five types of physiological signals: fNIRS, EEG, eye tracking, facial expression and ECG. Finally, HAT discrimination models with varying interpretability were developed using the identified features. Following this, six hypotheses were proposed:

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

Hypothesis 1 (H1): A significant difference in EEG α-band power exists between LHAT and HHAT levels.

Hypothesis 2 (H2): A significant difference in fNIRS HbO concentration exists between LHAT and HHAT levels.

Hypothesis 3 (H3): A significant difference in eye-tracking fixation metrics exists between LHAT and HHAT levels.

Hypothesis 4 (H4): A significant difference in facial micro-expression features exists between LHAT and HHAT levels.

Hypothesis 5 (H5): A significant difference in ECG-derived HRV exists between LHAT and HHAT levels.

Hypothesis 6 (H6): A high-interpretability discrimination model constructed based on HAT-sensitive physiological features can achieve an accuracy exceeding 80% in distinguishing between high and low HAT levels.

2.0 Methods

2.1 Experimental design

This study aims to explore physiological features sensitive to HAT and achieve effective discrimination. Given the theoretical ambiguity surrounding HAT's definition and formation, it is important to note that the focus of this study is not limited to HAT itself, but rather on the HAT-related state, which is manipulated through the factors of reliability and transparency. Throughout the manuscript, the term 'HAT state/level' refers specifically to this HAT-related state. In the experimental design, HAT level was treated as the independent variable, while the measured physiological features served as dependent variables. The HAT levels were operationalised through LHAT and HHAT scenario configurations in a simulated aviation monitoring and takeover task, which followed a 2 (reliability: low, high) $\times$ 2 (transparency: low, high) factorial design, resulting in four experimental conditions: low reliability–low transparency, low reliability–high transparency, high reliability–low transparency and high reliability–high transparency. The HAT levels were classified based on the Jian scale scores obtained from these four conditions. Furthermore, feature analysis was conducted on five categories of physiological measurements collected under different HAT levels: EEG, fNIRS, eye tracking, facial expression analysis and ECG.

2.2 Participants

This study recruited 24 participants (16 males, 8 females; mean age = 26.4 $\pm$ 5.13 years, range = 19–37) from the International Innovation Institute of Beihang University. All participants were right-handed, had normal or corrected-to-normal vision, were in good general health and had no history of neurological disorders. They were also required to have adequate sleep within 24 h prior to the experiment. The study strictly complied with the ethical principles of the Declaration of Helsinki (https://www.wma.net/policies-post/wma-declaration-of-helsinki/) and was approved by the Biomedical and Medical Ethics Committee of Beihang University (Approval No.: BM20240285). Before participation, all individuals were fully informed about the experimental procedures and precautions and provided written informed consent.

2.3 Experimental platform and tasks

HAT experiments were conducted in a simulated aviation operational environment using the Multi-Attribute Task Battery II (MATB-II) platform [37, 38]. The experimental stimuli were presented on a 27-inch AOC Q27G3Z/D monitor with a resolution of 1920 $\times$ 1080 pixels. Participants interacted with the system through a flight joystick and a mouse, and the specific configuration illustrated in Fig. 1.

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

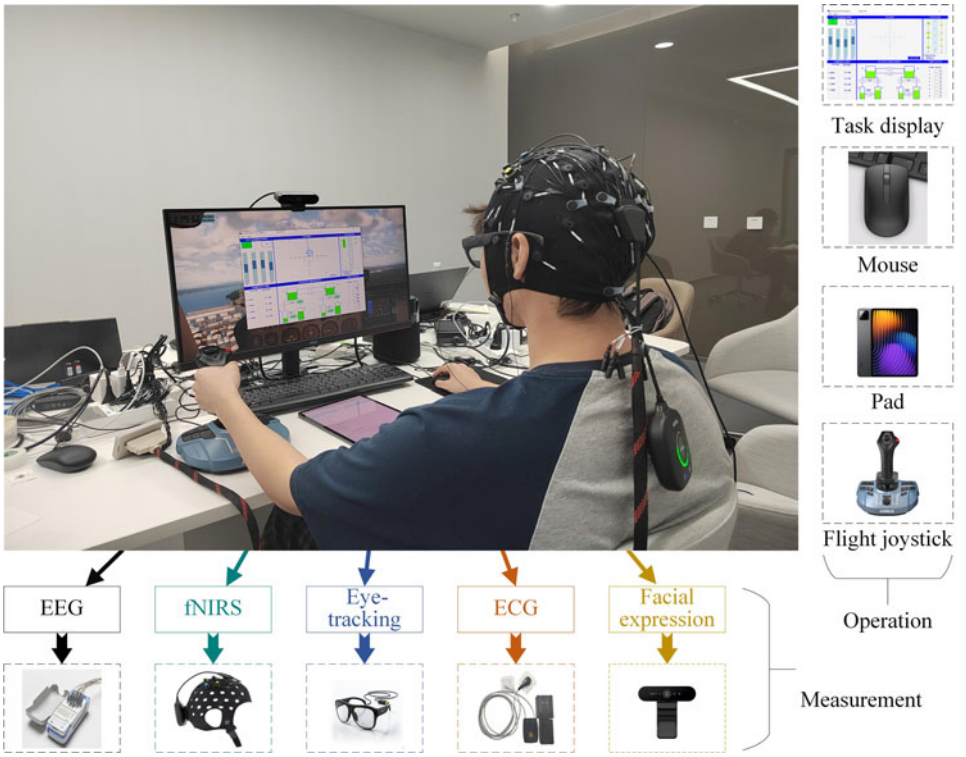


Figure 1. *Overview of operation and measurement devices in experimental platform.*

This experiment involved a simulated aviation monitoring and takeover task comprising three core subtasks: system monitoring (SYS), takeover and flight (TAK) and resource regulation and pump monitoring (RES), as illustrated in Fig. 2. In the SYS subtask, participants were required to continuously monitor the status of four gauges within the SYS area. Under normal conditions, the indicator values remained within an acceptable range; however, when they entered abnormal zones (the upper or lower three segments), participants were instructed to click the mouse to restore normal values. In the TAK subtask, the aircraft initially operated in autopilot mode. When specific conditions triggered autopilot disengagement, the takeover warning system (TWS) issued a five-second voice alert in advance, prompting participants to decide whether to initiate manual takeover by pressing the trigger button on the joystick, followed by adjusting the aircraft's attitude through manual control. The RES subtask required continuous monitoring of fuel consumption. When a pump malfunction occurred (indicated in red), participants were required to left-click the mouse to repair it while dynamically adjusting the on/off states of pumps 1–8 to maintain the fuel levels in main Tanks A and B within safe limits.

2.4 EEG equipment and related features

Physiological measurement devices include EEG, fNIRS, eye tracking, facial expression and ECG.

2.4.1 EEG devices and related features

EEG signals were recorded from 18 electrode sites (FP1, FP2, F7, F3, F4, FZ, F8, C3, C4, CZ, M1, M2, P7, P3, PZ, P4, P8, VEO) using a Neuroscan okti amplifier (Compumedics Ltd., Victoria, Australia). The system employed Ag/AgCl electrodes and was configured with a recording bandwidth of 0.1 to 200 Hz and a sampling rate of 1000 Hz, using GND as the ground electrode and REF as the online reference. EEG data processing comprised two main stages: preprocessing and analysis. The

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

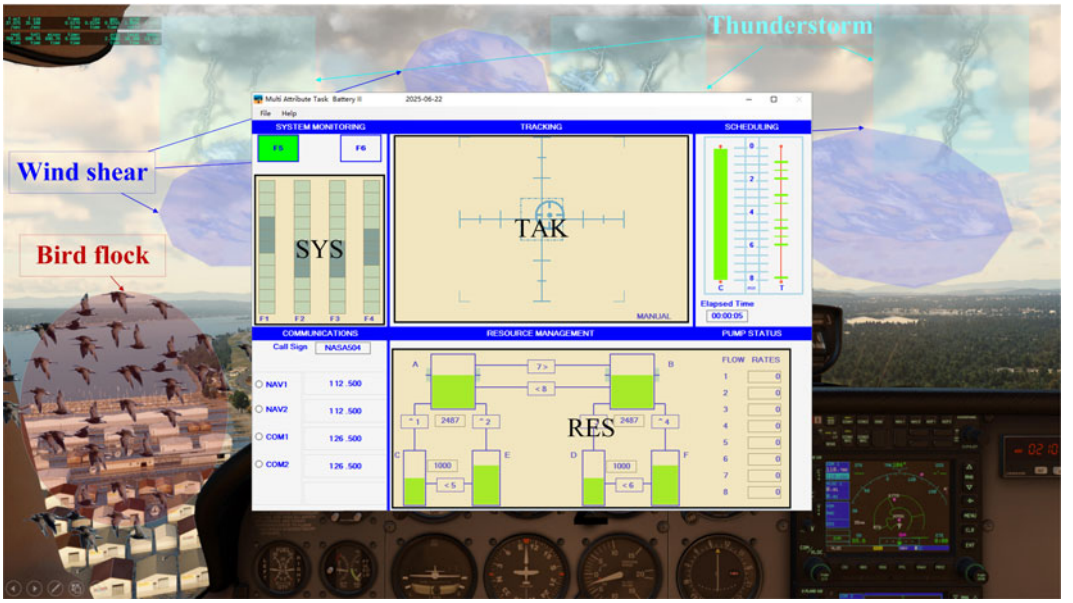


Figure 2. *Task display and subtask operation area (area of interest) layout.*

raw signals were first re-referenced to the average of the bilateral mastoid electrodes (M1 and M2). Subsequently, a 1–30 Hz bandpass filter was applied using the EEGLAB 2023.0 toolbox (available at https://sccn.ucsd.edu/eeglab/download/daily/) running on MATLAB R2020a. Independent component analysis (ICA), combined with the Adjust 1.1.1 toolbox (available at https://www.nitrc.org/frs/?group_id=739&release_id=3054#), was used to identify and remove artifacts such as eye movements [39]. The preprocessed signals were then segmented and subjected to fast Fourier transform (FFT). For frequency-domain analysis, absolute power (in μV^2) and relative power (%) [40] were computed for the δ (1–4 Hz), theta (θ) (4–8 Hz), α (8–12 Hz) and β (13–30 Hz) frequency bands using 15 electrode sites (excluding M1, M2 and VEO). Relative power for each band was defined as the ratio of its absolute power to the total absolute power across the 1–30 Hz range.

2.4.2 fNIRS devices and related features

Cerebral haemodynamic signals were acquired using the Artinis Brite Lite portable fNIRS system (Artinis Medical Systems B.V., Zetten, The Netherlands). The system was configured in a 2×4 eight-channel array comprising four source optodes and four detector optodes, with an inter-optode distance of 3.0 cm to ensure effective penetration of near-infrared light into cortical tissues. Utilising near-infrared spectroscopy at 760 nm and 850 nm, the system recorded changes in haemoglobin concentration at a 100 Hz sampling frequency. Data processing was conducted using the manufacturer's OxySoft analysis software (Artinis Medical Systems B.V., Zetten, The Netherlands), from which the concentration values of HbO and HbR (in mmol·mm) were extracted from eight predefined channels (R1_T1, R1_T3, R2_T1, R2_T3, R3_T4, R3_T5, R4_T4, R4_T5) for subsequent analysis as physiological features.

2.4.3 Eye-tracking devices and related features

Eye tracking data were recorded in real time using the Tobii Glass III wearable eye-tracking system (Tobii Technology AB, Stockholm, Sweden). The system operated at a sampling rate of 50 Hz with a spatial accuracy better than 0.5 degrees and offered a tracking range of 82 degrees horizontally and 48 degrees vertically. The scene camera resolution was set to 1920×1080 pixels with a frame rate of 25 Hz. A one-point calibration procedure was implemented, and adjustable lenses (ranging from 50 to 800 diopters)

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

were provided to participants requiring vision correction to ensure optimal tracking performance. As shown in Fig. 2, the experimental interface was divided into three areas of interest (TAK, RES and SYS). Eye-tracking features were extracted using Tobii Studio 3.4.5 software (Tobii Technology AB, Stockholm, Sweden) and categorised into three types: fixation features, which included the average duration of whole fixations (ADWF), average duration of fixations in the TAK area (ADF.TAK), total duration of fixations in the TAK area (TDF.TAK) and number of fixations in the TAK area (NF.TAK); saccade features, comprising the average amplitude of saccade (AAS), average peak velocity of saccades (APVS) and standard deviation of peak velocity of saccades (SDPVS); and pupillary features, represented by the average whole-fixation pupil diameter (AWFPD).

2.4.4 Facial expression devices and related features

Facial expressions of participants during the task were recorded and analysed using the FaceReader facial expression acquisition and analysis system (Noldus Information Technology, Wageningen, The Netherlands). The system was equipped with a scene camera configured at a resolution of 640×480 pixels and a frame rate of 30 fps, and it captured facial muscle Action Unit (AU) activity at a sampling frequency of 5 Hz. The extracted facial expression features comprised seven basic micro-expressions and 20 specific facial AUs. The seven micro-expressions were neutral, joy, sadness, anger, surprise, fear and disgust, with their mean values during the task used for analysis. The 20 selected AUs were as follows: inner brow raiser (AU01), outer brow raiser (AU02), brow lowerer (AU04), upper lid raiser (AU05), cheek raiser (AU06), lid tightener (AU07), nose wrinkler (AU09), upper lip raiser (AU10), lip corner puller (AU12), dimpler (AU14), lip corner depressor (AU15), chin raiser (AU17), lip pucker (AU18), lip stretcher (AU20), lip tightener (AU23), lip press (AU24), lips part (AU25), chin drop (AU26), mouth stretch (AU27) and eye closure (AU43).

2.4.5 ECG devices and related features

ECG signals were continuously recorded from participants using the BioBoX ECG acquisition system (Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Jiangsu, China) at a sampling rate of 250 Hz. Signal processing and feature extraction were conducted using the SIBET analysis software developed by the same institute. The time-domain features included beats per minute (BPM) and SDNN, while frequency-domain features consisted of low frequency power (LF), total power (TP), low frequency ratio (LF ratio), high frequency ratio (HF ratio), low frequency in normalised units (LF_nu) and the ratio of low frequency to high frequency (LF/HF).

2.5 Procedures

The total duration of the experiment was approximately 120 min, divided into two phases: training and formal testing. The training phase, lasting about 40 min, involved signing informed consent forms, completing a demographic questionnaire, practicing the simulated aviation monitoring and takeover task (participation in the formal experiment was contingent upon passing an assessment) and calibrating the five physiological measurement devices. The formal experiment phase lasted approximately 80 min, beginning with a 1-min resting-state baseline test, followed by four experimental scenarios with varying levels of reliability and transparency. After completing each scenario, participants filled out the Jian scale and took a brief rest to ensure data quality and maintain engagement.

3.0 Results and analysis

Figure 3 presents the three main modules of the results and analysis, outlining the implementation steps for each module. The top layer shows the grouping of HAT levels, progressing from left to right through the development of the platform and tasks to complete the HAT level grouping. The middle layer focuses on the sensitive physiological feature analysis across different HAT levels, with the analysis progressing

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

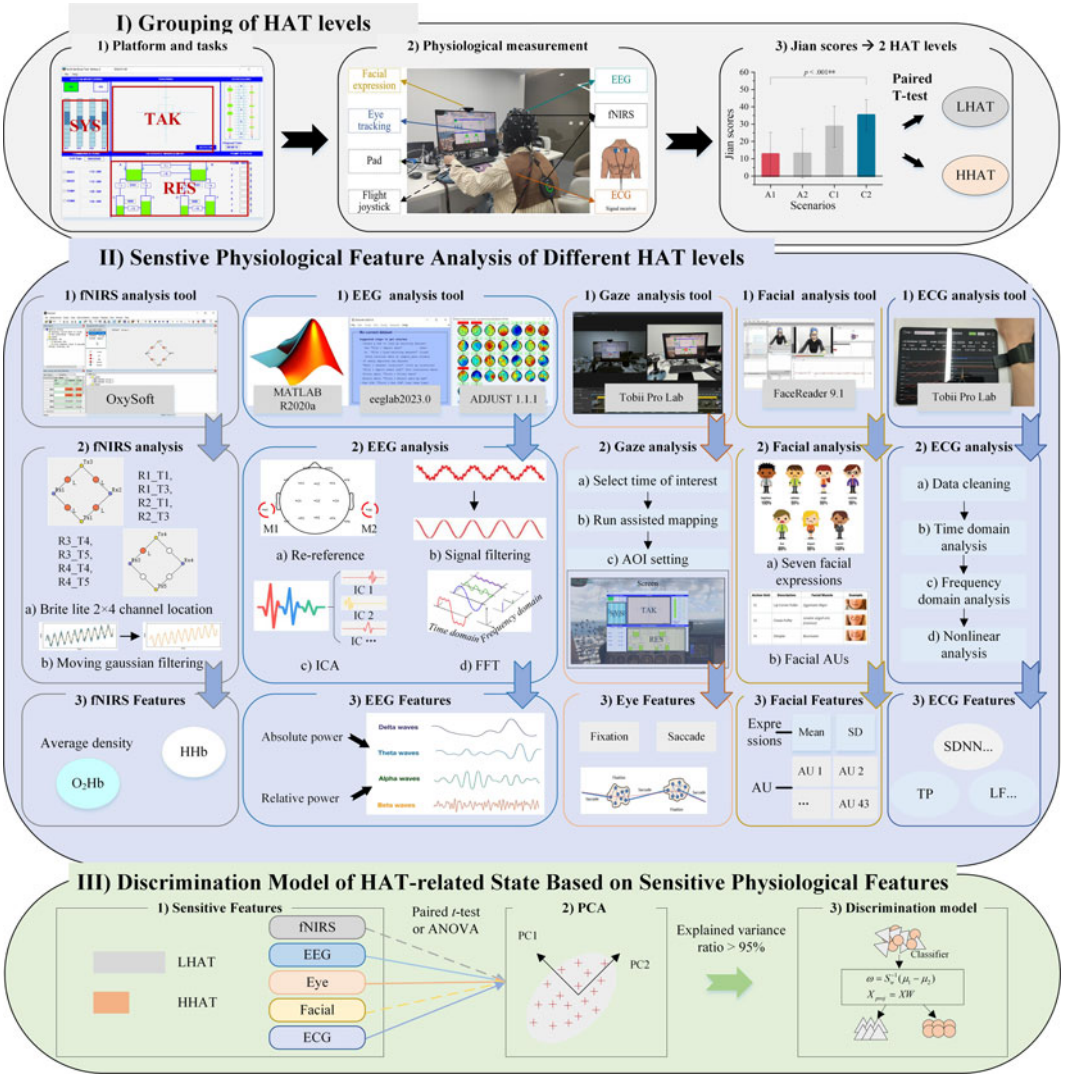


Figure 3. *Overview of data analysis methods and physiological features.*

from top to bottom. The bottom layer illustrates the discrimination model of HAT-related states based on sensitive physiological features, where HAT-related state discrimination is achieved by selecting sensitive features, progressing from left to right.

3.1 Grouping of HAT statues

Considering the validity of the Jian scale in assessing HAT [5, 41], this study classified scenarios into LHAT and HHAT conditions based on their Jian scale scores, as illustrated in Fig. 3. Descriptive statistics revealed a progressive increase in Jian scores across the four experimental scenarios: low reliability–low transparency (12.65 ± 12.53), low reliability–high transparency (13.08 ± 14.20), high reliability–low transparency (28.63 ± 11.92) and high reliability–high transparency (35.21 ± 8.97). Current definitions of HAT remain inconsistent, with over 50 distinct definitions reported in the literature [7, 42]. Previous research has identified system transparency and reliability as key predictors of HAT [16, 43], with HAT levels generally increasing as these factors improve. In addition, several models

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

and frameworks have been proposed by researchers to explain the formation mechanisms of HAT, incorporating these two important factors [44, 45]. Building on these prior frameworks, the present study adopts an operational definition of HAT levels grounded in transparency and reliability and empirically validates this definition through measurement. To minimise label noise, the low reliability–low transparency scenario, which yielded the lowest Jian scores, was categorised as the LHAT condition, while the high reliability–high transparency scenario, associated with the highest scores, was classified as the HHAT condition. A paired samples t-test confirmed that Jian scores under the LHAT condition were significantly lower than those under the HHAT condition ($t(23) = -9.067$, $p < 0.001$). These results collectively support the validity of the two-scenario HAT grouping approach adopted in this study.

Furthermore, no significant differences were observed between LHAT and HHAT conditions in attentional demand scores (3D SART scale [46]) or mental workload scores on the 0–10 scale (LHAT: 5.03 ± 2.19, HHAT: 4.58 ± 2.67, $t(23) = 1.359$, $p > 0.05$; LHAT: 4.43 ± 1.47, HHAT: 4.00 ± 1.82, $t(23) = 1.64$, $p > 0.05$), suggesting that, within the present design, attentional demand and mental workload do not explain the observed physiological differences between HAT conditions. In a subsequent analysis, a 2 (reliability) × 2 (transparency) two-way repeated-measures analysis of variance (ANOVA) was conducted. The analysis revealed no significant reliability × transparency interaction effects on the total scores of the Jian scale ($p > 0.05$). Further, a significant main effect of reliability was observed, with significantly higher trust scores under high-reliability conditions, $F(1, 22) = 55.733$, $p < 0.001$, $\eta_p^2 = 0.717$. The main effect of transparency was marginally significant, with trust scores tending to be higher under high-transparency conditions, $F(1, 22) = 4.132$, $p = 0.054$, $\eta_p^2 = 0.158$. These findings provide insights into how transparency and reliability influence HAT levels in this experiment. The decision to collapse the conditions into LHAT and HHAT was made to facilitate physiological state differentiation while ensuring alignment with statistically validated trust drivers. This approach prioritises HAT-related state differentiation, while preserving the underlying experimental manipulations of reliability and transparency.

3.2 Sensitive physiological feature analysis of different HAT levels

This section conducts a sensitive physiological feature analysis across different modalities, aiming to validate the physiological features associated with H1–H4 and to further explore additional candidate indicators. For EEG data, electrode placements in two brain regions (central: C3, Cz, C4; parietal: P3, Pz, P4), which are closely associated with HAT-related states, were selected for analysis [8]. In addition, laterality is commonly used to examine differences among left, middle and right hemispheric activities [47, 48]. Accordingly, three lateralities were defined: left (C3, P3), middle (Cz, Pz) and right (C4, P4). Based on this framework, the absolute and relative power of four EEG frequency bands (δ, θ, α and β) were analysed using a 2 (HAT condition: LHAT vs. HHAT) × 2 (brain region: central vs. parietal) × 3 (laterality: left, middle, right) repeated-measures ANOVA. Mauchly's test was used to assess sphericity, and when the assumption was violated, the Greenhouse–Geisser correction was applied. Post hoc multiple comparisons were adjusted using the Bonferroni method [49]. For fNIRS, eye-tracking, facial expression and ECG features, paired-sample t-tests were conducted for the hypothesis-driven (confirmatory) features related to H2–H4. For exploratory features, those showing significant differences in the paired t-tests were further subjected to Bonferroni-corrected multiple comparisons to control for Type I error.

3.2.1 Analysis of EEG features

Figure 4 presents topographic maps displaying the descriptive visualisation of absolute power across four EEG bands for all participants. Warmer colours (red) indicate higher activation at corresponding electrode sites, while cooler colours (blue) represent lower activation. Repeated-measures ANOVA results showed no significant interaction effects of absolute power ($ps > .05$). Moreover, the main effect of the HAT condition on absolute power in four waves was also not significant ($ps > .05$).

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

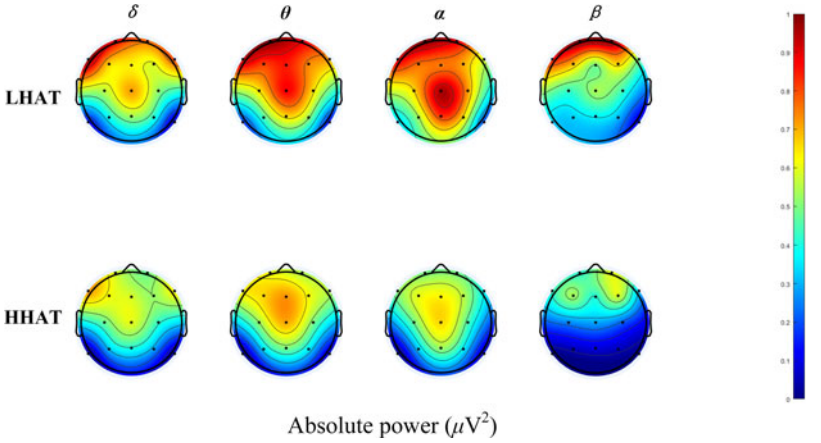


Figure 4. *Descriptive visualisation of four EEG absolute power under different HAT levels.*

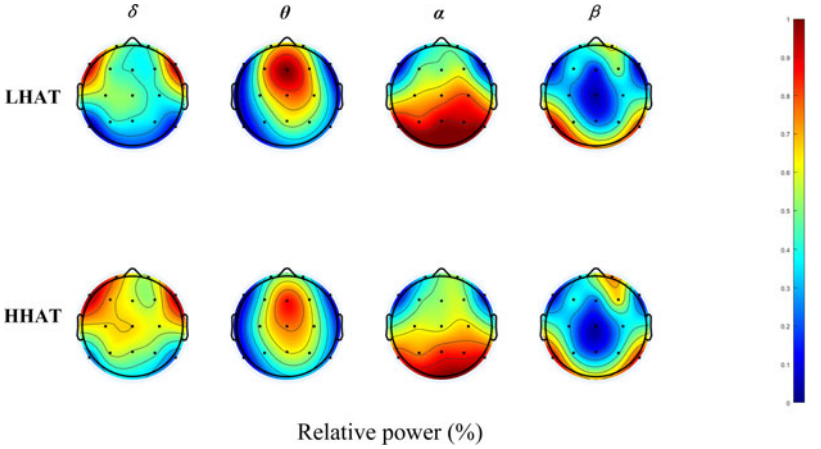


Figure 5. *Descriptive visualisation of four EEG relative power under different HAT levels.*

As illustrated in Fig. 5, the EEG topographic maps provide a descriptive visualisation of relative spectral power for the four bands. As for δ relative power, the main effect of the HAT condition is also significant; the LHAT condition shows significantly lower δ than the HHAT condition ($p = 0.020$). As for α relative power, there are significant interaction effects found in the HAT condition $\times$ region $\times$ laterality ($F(1.566, 36.027) = 4.153$, $p = 0.032$, $\eta^2 = 0.153$). Simple effect analysis showed that in the central brain region, the α relative power of the HAT condition is significantly higher than the HHAT condition in midline ($p = 0.048$), and right laterality ($p = 0.009$); the main effect of the HAT condition is also significant, the LHAT condition shows significantly higher α than the HHAT condition ($p = 0.033$). The results are shown in Table 1.

3.2.2 Analysis of fNIRS features

Figure 6 illustrates the descriptive results of HbO and HbR across eight channels under different HAT levels. Subsequent paired *t*-tests revealed no statistically significant differences in these fNIRS features between high and low HAT levels ($ps > 0.05$).

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

Table 1. *ANOVA results of relative power under different HAT levels*

Waves	Region × laterality × HAT condition			Region × HAT condition			Laterality × HAT condition			HAT condition		
(%)	F	p	η^2	F	p	η^2	F	p	η^2	F	p	η^2
δ	3.047	0.077	0.117	0.333	0.569	0.014	0.130	0.753	0.006	6.246	0.020*	0.214
θ	0.232	0.712	0.010	3.203	0.087	0.122	0.410	0.547	0.018	0.162	0.691	0.007
α	4.153	0.032*	0.153	0.278	0.603	0.278	0.678	0.479	0.029	5.173	0.033*	0.184
β	3.319	0.073	0.126	2.425	0.133	0.095	1.955	0.166	0.078	1.864	0.185	0.075

Note: *indicate $p < 0.05$.

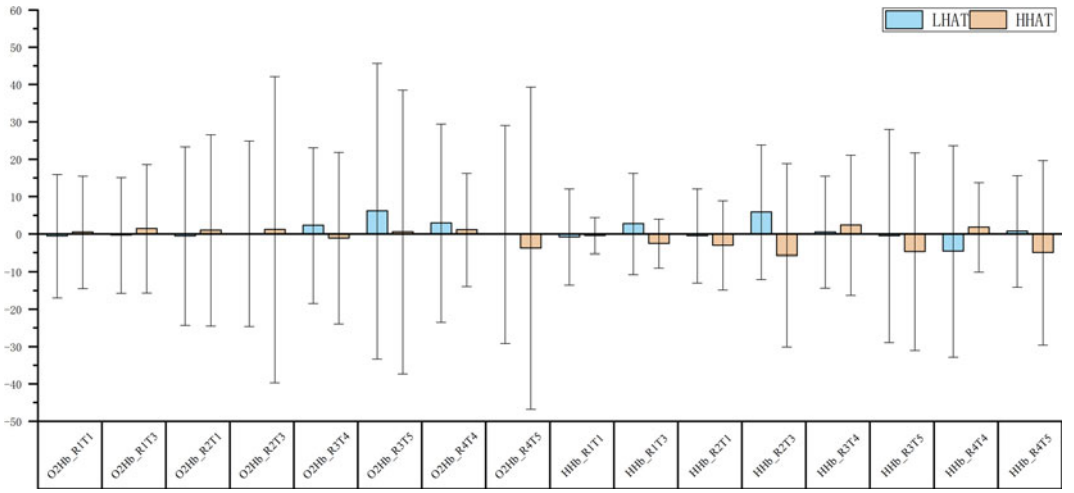


Figure 6. *fNIRS features under different HAT levels. Error bars represent standard deviation (SD); the same convention applies to the subsequent figures.*

3.2.3 Analysis of eye-tracking features

Descriptive results for eye-tracking features, specifically fixation, saccade and pupillary response, are presented in Fig. 7.

Paired t-tests revealed that, compared to the HHAT condition, the LHAT condition was associated with significantly shorter durations in the following features: TDF.TAK, ADF.TAK and ADWF ($ps < 0.05$), as detailed in Fig. 7 and Table 2. Furthermore, for the exploratory features not involved in H3 (i.e. AAS, APVS, SDPVS and AWFPD), Bonferroni-corrected comparisons indicated no significant differences ($ps >$ the adjusted threshold of 0.0125).

3.2.4 Analysis of facial expression features

Descriptive results for the seven categories of micro-expressions and 20 AUs are presented in Figs. 8 and 9. Subsequent paired t-tests indicated no statistically significant differences in these facial expression features across different HAT levels ($ps > 0.05$).

3.2.5 Analysis of ECG features

Descriptive results of ECG features under different HAT levels are presented in Fig. 10.

Paired t-tests revealed that, compared to the HHAT condition, the LHAT condition showed significantly increased values in SDNN, LF power, TP, LF ratio, LF_nu and LF/HF ($ps < 0.05$), as summarised

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

Table 2. *Paired t-test results of typical fixation features under different HAT levels*

Fixation features of eye-tracking	LHAT ($M \pm SD$)	HHAT ($M \pm SD$)	t	df	p
TDF.TAK * 10^{-3} (ms)	119.363 ± 46.311	157.055 ± 47.168	-4.84	23	$< 0.001^{**}$
NF.TAK (N)	280.79 ± 129.23	325.21 ± 198.75	-1.53	23	0.140
ADF.TAK (ms)	471.88 ± 249.75	623.17 ± 362.54	-3.95	23	$< 0.001^{**}$
ADWF (ms)	427.00 ± 180.99	475.33 ± 183.85	-2.59	23	$< 0.05^{*}$

Note: **indicate $p < 0.001$; *indicate $p < 0.05$.

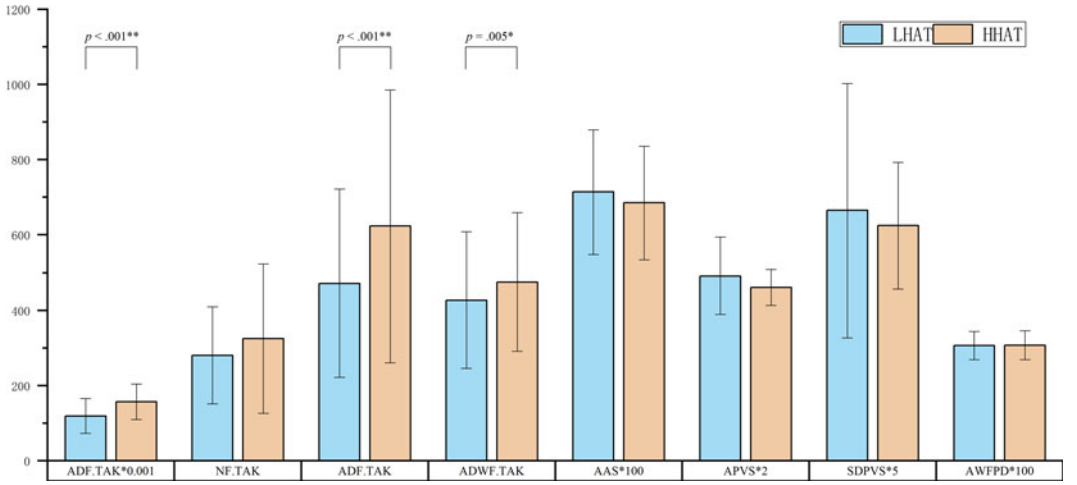


Figure 7. *Descriptive results of eye-tracking features under different HAT levels. The p-values and asterisks (*) indicate features showing significant differences in paired t-tests. The same convention applies to the subsequent figures.*

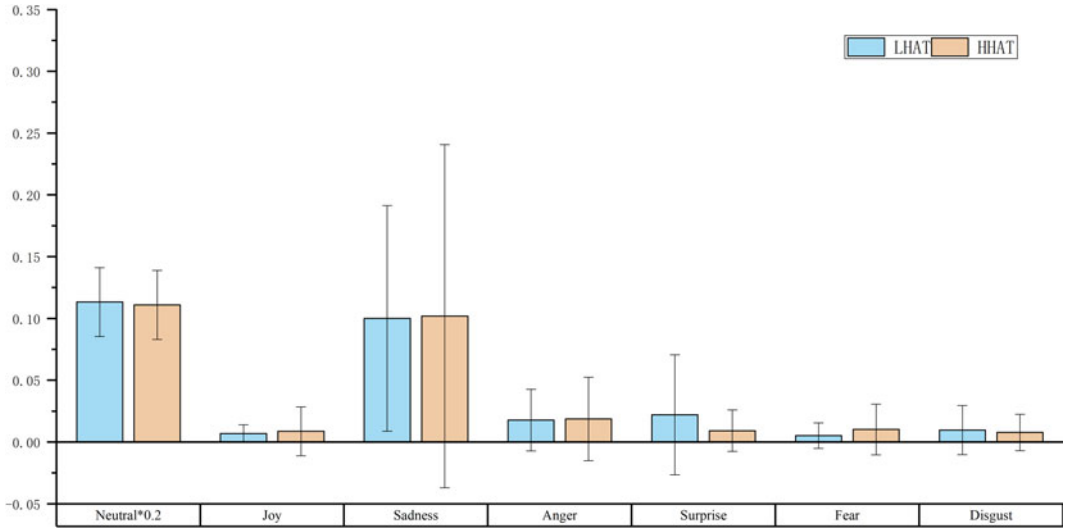


Figure 8. *Descriptive results of the mean values of micro-expressions under different HAT levels.*

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

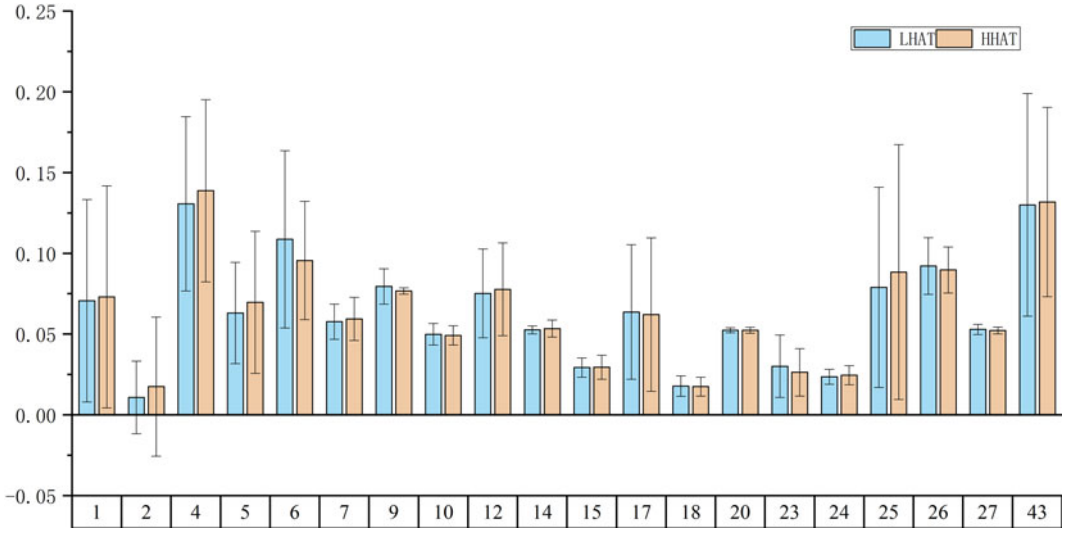


Figure 9. *Descriptive results of AUs under different HAT levels.*

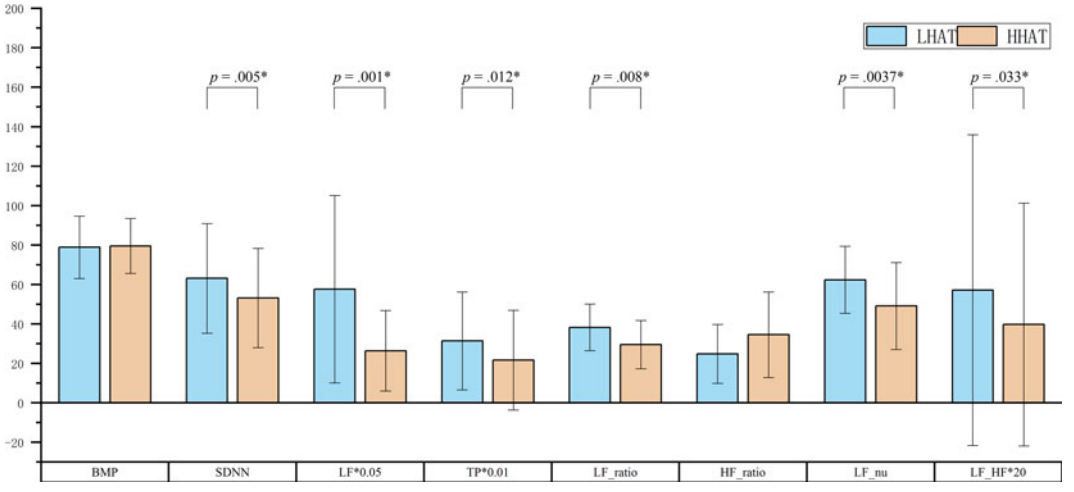


Figure 10. *Descriptive results of ECG features under different HAT levels.*

in Table 3. Due to electromagnetic interference and electrode gel conductivity issues, eight ECG data points were abnormal or missing, reducing degrees of freedom. For the exploratory features not involved in H4 (i.e. BPM), Bonferroni-corrected comparisons indicated no significant differences ($p > 0.05$).

3.3 Discrimination model of HAT based on sensitive physiological features

Physiological features showing significant differences between LHAT and HHAT levels in paired t-tests or repeated-measures ANOVA were identified as sensitive features as candidate HAT-sensitive features for interpretative analysis. Based on these analyses, 21 statistically significant features were selected as model inputs (see Table 4). These features were identified prior to model training, were not selected within the cross-validation procedure and remained fixed throughout the entire modeling process as inputs to supervised learning algorithms.

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

Table 3. *Paired t-test results of typical HRV features under different HAT levels*

ECG HRV features	LHAT ($M \pm SD$)	HHAT ($M \pm SD$)	t	df	p
SDNN (ms)	63.12 ± 27.80	53.14 ± 25.13	3.220	17	0.005*
LF (ms^2)	1151.51 ± 950.13	527.61 ± 408.23	3.9316	17	0.001*
TP (ms^2)	3137.74 ± 2479.52	2165.64 ± 2525.65	2.803	17	0.012*
LF ratio (%)	38.27 ± 11.85	29.55 ± 12.23	2.976	17	0.008*
HF ratio (%)	24.83 ± 14.92	34.53 ± 21.73	−1.793	17	0.091
LF_nu	62.36 ± 16.98	49.12 ± 22.01	3.361	17	0.004*
LF/HF	2.86 ± 3.94	1.99 ± 3.08	2.327	17	0.033*

Note: *indicate $p < 0.05$.

Table 4. *Input features for supervised learning models*

Number	Category of model input	Model input
1∼3	Eye tracking	ADF.TAK, TDF.TAK
		ADWF
4∼9	ECG	SDNN
		LF ratio, TP, LF/HF, LF, LF_nu
10∼21	EEG	δ_r_C3/CZ/C4/P3/PZ/P4
		α_r_C3/CZ/C4/P3/PZ/P4

Note: ** and * indicate $p < 0.001$ and $p < 0.05$, respectively.

The data preprocessing pipeline included z-score standardisation and principal component analysis (PCA), retaining components that accounted for over 95% of the cumulative variance. All preprocessing steps were strictly implemented within a nested cross-validation framework using training folds only via a pipeline structure, ensuring that no information from the test data was used during model training. The selected features were treated as fixed inputs and were not re-selected, re-weighted or updated within any cross-validation folds. Therefore, no information from the test folds was used in model training, and no data leakage occurred during the machine learning process. Subsequently, nine classifiers from the Python *sklearn* library were employed for trust-sensitive discrimination: decision tree (DT), Gaussian naive Bayes (GNB), LR, LDA, KNN, multilayer perceptron (MLP), RF and support vector machine (SVM) with both linear and nonlinear kernels [50]. The first group of classifiers (DT, GNB, LR, LDA, linear SVM) were selected for their high interpretability, whereas the latter four classifiers (KNN, MLP, RF, nonlinear SVM) were characterised by lower interpretability. The model training and evaluation protocol was implemented using a nested cross-validation framework. Specifically, (1) optimal hyperparameters were identified via inner five-fold cross-validation; (2) classification performance was evaluated on outer test folds (five-fold, eight-fold and ten-fold) using accuracy, precision, recall and F1-score; (3) performance metrics were averaged across folds and (4) the entire procedure was repeated 20 times with different random seeds to obtain robust estimates. As summarised in Table 5, the interpretable linear SVM classifier achieved the best HAT discrimination performance, with an accuracy of 80.65% and an F1-score of 79.42%.

4.0 Discussion

To identify HAT-sensitive physiological features and accurately discriminate between high and low HAT levels, this study designed a simulated aviation monitoring and takeover task based on a 2 (reliability) × 2 (transparency) factorial design. High and low HAT levels were classified according to their Jian scale scores under different task scenarios, and paired t-tests confirmed significantly lower scores in the LHAT condition than in the HHAT condition. Statistical analyses of five categories of physiological

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

Table 5. *The discriminative performance of typical classifiers*

Classifiers	Outer folds	Inner folds	Mean ± standard deviation ($M \pm SD$, %)			
			Accuracy	Precision	Recall	F1 Score
DT	5	5	61.83 ± 6.13	63.18 ± 7.50	61.82 ± 6.03	60.46 ± 6.41
GNB			64.42 ± 4.55	65.62 ± 5.10	64.57 ± 4.52	63.70 ± 4.59
LR			75.01 ± 2.77	77.11 ± 3.21	75.20 ± 2.89	74.38 ± 2.83
LDA			76.99 ± 3.34	78.94 ± 3.71	77.10 ± 3.32	76.45 ± 3.53
Linear SVM			78.90 ± 3.08	81.19 ± 3.58	79.10 ± 3.14	78.34 ± 3.15
KNN			59.54 ± 7.42	59.85 ± 8.43	59.58 ± 7.51	58.46 ± 7.69
MLP			69.82 ± 3.61	72.10 ± 4.54	70.03 ± 3.71	68.87 ± 3.69
RF			61.91 ± 5.97	63.35 ± 6.19	62.10 ± 5.93	61.09 ± 6.04
Nonlinear SVM			65.91 ± 6.26	68.58 ± 7.47	66.23 ± 6.37	64.64 ± 6.50
DT	8	5	59.58 ± 7.04	60.34 ± 8.98	59.58 ± 7.04	57.92 ± 7.54
GNB			64.27 ± 2.89	65.23 ± 4.44	64.27 ± 2.89	62.44 ± 3.37
LR			76.88 ± 3.70	79.60 ± 4.30	76.88 ± 3.70	75.93 ± 3.79
LDA			75.62 ± 3.65	79.03 ± 4.40	75.62 ± 3.65	74.37 ± 4.01
Linear SVM			80.52 ± 4.65	83.02 ± 5.22	80.52 ± 4.65	79.81 ± 4.83
KNN			63.02 ± 5.32	64.40 ± 6.02	63.02 ± 5.32	61.32 ± 5.52
MLP			68.23 ± 4.32	70.65 ± 4.85	68.23 ± 4.32	66.33 ± 4.35
RF			63.23 ± 4.29	64.78 ± 6.28	63.23 ± 4.29	61.62 ± 4.59
Nonlinear SVM			64.69 ± 5.99	66.46 ± 7.58	64.69 ± 5.99	62.71 ± 5.88
DT	10	5	63.22 ± 6.09	61.82 ± 8.37	62.79 ± 6.44	59.67 ± 6.87
GNB			65.25 ± 4.06	67.12 ± 5.22	65.67 ± 4.30	63.00 ± 4.47
LR			77.25 ± 3.19	79.25 ± 4.43	77.04 ± 3.16	75.53 ± 4.00
LDA			75.85 ± 3.75	78.16 ± 4.84	75.83 ± 3.83	74.06 ± 4.25
Linear SVM			**80.65 ± 2.75**	**83.00 ± 3.18**	**80.67 ± 2.84**	**79.42 ± 2.96**
KNN			60.08 ± 5.07	60.09 ± 6.31	60.33 ± 5.44	57.31 ± 6.03
MLP			68.53 ± 4.54	69.82 ± 4.86	68.54 ± 4.49	66.15 ± 4.59
RF			61.05 ± 5.89	60.17 ± 7.74	60.96 ± 6.19	57.91 ± 6.24
Nonlinear SVM			66.10 ± 5.88	65.91 ± 7.13	66.71 ± 5.55	63.13 ± 6.10

signals, including fNIRS, EEG, eye tracking, facial expression and ECG, revealed that the LHAT condition, relative to HHAT, was associated with significant increases in EEG α relative power, decreases in δ relative power, shorter eye-tracking metrics such as ADWF, ADF.TAK and TDF.TAK and elevated ECG metrics including SDNN, LF/HF and LF ratio. Using these HAT-sensitive physiological features, a highly interpretable linear SVM model was constructed for HAT classification, achieving a discrimination accuracy of 80.65% and an F1-score of 79.42%. These results demonstrate that the fusion of multimodal physiological features can effectively support the objective identification of HAT levels in aviation environments.

Experimental results demonstrated that EEG α relative power under LHAT was significantly higher than under HHAT, thereby supporting H1. This finding aligns with results reported by Blais et al.in a coin toss game [14], but contrasts with observations in simulated automated driving tasks [12], decision-making scenarios involving automatic versus manual control [10], word elicitation studies [11] and scaffolding design and rectification tasks [13]. Elevated α activity is generally associated with a state of relaxed alertness and the active suppression of task-irrelevant external stimuli [51, 52]. Under distrust conditions, however, the brain may recruit additional cognitive resources to process uncertainty, suggesting that increased α power reflects efforts to maintain focus on primary tasks while suppressing interference induced by distrust. Additionally, δ relative power was significantly lower under LHAT compared to HHAT. δ waves are thought to be associated with deep relaxation or unconscious states

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

[53], whereas distrust might involve heightened alertness or anxiety, which could activate stress-related neural pathways and potentially suppress δ activity. Although these findings suggest that both α and δ relative power are sensitive to variations in HAT-related status, previous research has also highlighted that α and δ reflect multiple overlapping processes, including workload, vigilance, attentional disengagement and inhibitory control. Therefore, the relationship between these measures and HAT remains to be further explored. At the regional level, central electrodes (C3/CZ/C4), which are associated with emotional processing and conflict monitoring, and parietal sites (P3/PZ/P4), which are involved in spatial attention and sensory integration, were implicated in HAT-related cognitive processes, providing neural insights into the underlying mechanisms. In contrast, no significant differences in fNIRS HbR or HbO concentrations were observed between LHAT and HHAT, resulting in the rejection of H2. This result diverges from findings in non-driving tasks during highly automated driving [15], crime mapping tasks [16], image classification tasks with an AI teammate [17] and collaborative tasks with a UR10 robot [18]. These discrepancies may be attributed to contextual differences across experimental setups; nevertheless, the present findings offer new insights into trust assessment from an fNIRS perspective.

Additional analyses were conducted on eye tracking, facial expression and ECG signals under high and low HAT levels. Supporting H3, eye-tracking metrics, including ADWF, ADF.TAK and TDF.TAK, exhibited significant sensitivity to different HAT levels, with all three fixation metrics being significantly shorter under LHAT compared to HHAT. This pattern may stem from operators frequently shifting visual attention between critical takeover areas and other monitoring points under distrust conditions, coupled with the exclusively auditory alerts delivered by the TWS in this experiment, which distracted their visual attention. Although these findings differ from several previous studies reporting increased fixation duration and frequency under distrust [29, 54, 55], as well as from results of a UAV target identification task [20] where fixation behaviour showed no significant variation across reliability levels, they provide new and actionable eye-tracking features for HAT classification in comparable scenarios. In contrast, no significant differences in facial micro-expressions or AUs were detected across HAT levels, leading to rejection of H4. This outcome contradicts the expectation that affective trust breaches often trigger observable negative emotional expressions [54] and differs from findings in a bluff game task study [23], likely due to the formal and high-stakes context of aviation tasks suppressing natural emotional expressivity. For ECG measures, significant increases in SDNN, LF/HF and LF ratio were observed under LHAT relative to HHAT, supporting the acceptance of H5. This finding contrasts with results from a planetary gear assembly task [21]. A plausible physiological explanation is that distrust triggers a state of alertness or tension, activating the sympathetic nervous system (reflected by elevated LF power and LF/HF ratio) while suppressing parasympathetic activity, leading to an overall increase in HRV as indexed by SDNN.

Based on paired *t*-test results and repeated-measures ANOVA, this study identified physiological features sensitive to HAT. After PCA-based dimensionality reduction, these features were used as input to a supervised machine learning model. Results showed that the linear SVM based on the selected sensitive features achieved an accuracy exceeding 0.80 under nested cross-validation (outer eight-fold or five-fold, inner five-fold), thereby supporting H6. Table 6 summarises the validation outcomes for H6 and related hypotheses, along with comparative results from previous studies and the experimental results. The discriminative performance attained in this study can be attributed to the rigorous selection of sensitive features and the consideration of their relative contributions. By excluding conflicting features during feature selection, the risk of classification confusion was effectively reduced. In terms of algorithm selection, this study differs notably from existing HAT models. Whereas prior studies often employed less interpretable models, such as RF or CNN [29, 31, 32], the present results demonstrate that an linear SVM model, characterised by higher interpretability, can also achieve strong performance when supplied with well-selected HAT-sensitive physiological features. This finding offers a feasible direction for implementing trust discrimination in real-world scenarios.

By analysing physiological features across different HAT levels and distinguishing between them, this study aims to elucidate the underlying mechanisms of trust degradation and propose feasible

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

Table 6. *Summary of assumptions and results*

Categories	EEG	fNIRS	Eye tracking	Facial expression	ECG	Discrimination model (number of categories, classifier, accuracy)
Previous research (LHAT)	1) α, $\beta\downarrow$; $\gamma\uparrow$ [10, 11]; 2) $\beta\uparrow$ (T. [8]; 3) $\alpha\uparrow$ [14]; 4) $\alpha\downarrow$ [13]; 5) δ, γ [9]; 6) $\alpha\downarrow$ [12].	1) HbO↑ [15]; 2) HbO↑; HBR nonsignificant [16]; 3) HbO & HBR↓ [17]; 4) HbO↑ [18].	1) Low automation trust, monitoring frequency of the automation ↑ [56]; 2) For low-reliability UAVs, total fixation duration, gaze fixations, scan paths↑, average fixation duration nonsignificant [20]; 3) Fixation count↑, gaze transition entropy↓ [21]; 4) Robot unreliability, more fixations, stationary gaze entropy↑; gaze transition entropy↓ [22].	1) Happy, smile↓ [23].	1) LF/HF↓ [25]; 2) HR, SDNN, LF/HF were all non-significant (Y. [21]; 3) HR↓ [27].	1) Three trust levels (lower, medium, higher), RF, 79.57% [28]; 2) Two trust-change directions (trust increased vs trust decreased), RF, 73.7% [29]; CNN-LSTM [30]; 3) Three trust levels (low, medium, high), LightGBM, 88.44%(T. [8]; 4) Two classes (trust vs mistrust), XGBoost, 76.0% [17]; 5) Changes in trust (increasing or decreasing trust), CNN, 71.9% [31]; 6) Two trust levels (high *vs* low), CNN, 78.61% (handing task), 73.35% (pouring task) [32]; 7) Two classes (trust vs distrust), XGBoost, 81.6% [24]; 8) Ten trust levels (classes 0 through 9), CNN, 94.3% [33]; 9) Two classes (Trust vs distrust), QDA, 78.58% [34]; 10) Two trust levels (high vs low), RF, 92.97% (individual); 81.51% (group + eye metric); 80.72% (group + raw eye data) [35].
Hypothesis	α	HbO	Fixations	Micro-expression	HRV	Accuracy > 80%
Results (LHAT)	$\delta\downarrow$, $\alpha\uparrow$	Nonsignificant	ADWF, ADF.TAK, TDF.TAK↓	Nonsignificant	SDNN, LF, LF/HF↑	Two HAT levels (high *vs* low), 80.65%
Validation of H1∼H6	✓	×	✓	×	✓	✓

prevention and alerting strategies to ensure the safe and efficient accomplishment of aviation-related tasks. The distinctive contributions of this research are presented in three aspects of an explainable 'output–input–discrimination algorithm' framework. First, regarding output interpretability, in contrast to previous studies that rarely focused on aviation-specific contexts, this study designed a simulated aviation monitoring and takeover task that incorporated both reliability and transparency considerations. Based on the Jian scale, two distinct HAT levels (HHAT and LHAT) were clearly defined, providing an interpretable labeled basis for subsequent exploration of sensitive features and discrimination. Second, with respect to input interpretability and the exploration of HAT-sensitive features, while existing studies typically limited physiological assessments to three categories [57] and often failed to clarify the relationship between physiological responses and HAT, this study innovatively employed five physiological categories. The findings revealed that EEG α and δ relative power, eye-tracking metrics (such as ADWF, ADF.TAK and TDF.TAK) and ECG metrics (such as SDNN, LF/HF and LF ratio) were sensitive to variations in HAT levels, offering explainable features for a discriminant classifier. Finally, concerning algorithmic interpretability, to address the common limitations of restricted interpretability and prediction accuracy in existing HAT discrimination methods, this study integrated the sensitive features into an LDA model, achieving a discriminative outcome with both high interpretability and accuracy.

However, this study has several limitations. First, as a pilot investigation conducted in a laboratory environment, its findings require further validation in real-world task scenarios, particularly with professional pilots training in high-fidelity flight simulators. Second, given the complexity of defining HAT and its underlying drivers, research should not only focus on HAT itself but also consider the HAT-related state. This study primarily explored cross-modal features of HAT-related states manipulated by reliability and transparency and identified their dynamics. These insights are crucial for refining the operational definition of HAT and advancing both theoretical and empirical research. Future research could explore HAT manipulation more closely aligned with real flight scenarios, defining three or more HAT levels to better uncover the underlying mechanisms linking physiological features to trust dynamics. Furthermore, it should be noted that manipulations of reliability and transparency may still induce variations in other task-related factors, such as workload, monitoring demand and takeover frequency, which are inherent to the characteristics of HAT and may be associated with attentional and cognitive states. Although we measured mental workload and attentional demand in the current experiment to address these potential confounds, future studies should continue to optimise experimental paradigms to more clearly disentangle HAT-related states from other task-related processes. Deeper exploration of HAT-sensitive features, alongside the incorporation of additional labeled data, would facilitate more interpretable trust assessment. Although this study verified the feasibility of using physiological features to evaluate and distinguish HAT levels, future work could integrate other physiological measures, such as GSR and functional magnetic resonance imaging (fMRI), to extend and validate the findings. Additionally, research could investigate optimal combinations of physiological sensors for flight environments, minimal configuration requirements and suitable non-contact approaches, thereby facilitating the translation of laboratory findings to real-time monitoring in operational settings. Finally, subsequent studies should examine different participant groups, including male and female cohorts, as well as populations such as professional pilots, flight cadets and students with an aviation background, to explore potential group differences in physiological features of HAT, while also expanding the sample size and broadening the range of measured physiological features.

5.0 Conclusions

This study classified LHAT and HHAT levels based on Jian scores during a simulated aviation monitoring and takeover task and analysed associated physiological signals to identify HAT-sensitive features across five categories: EEG, fNIRS, eye tracking, facial expression and ECG. A machine learning model incorporating these features was subsequently developed to discriminate between HAT levels. The results indicated that: (1) compared with HHAT, LHAT showed significantly decreased δ relative

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

power and increased α relative power in EEG; (2) no significant differences were found in fNIRS HbR and HbO concentrations between HAT levels; (3) the LHAT exhibited significantly shorter durations in eye-tracking metrics, including ADWF, ADF.TAK and TDF.TAK; (4) none of the seven micro-expressions derived from facial analysis showed significant differences across different HAT levels; (5) the highly interpretable linear SVM model integrating these sensitive physiological features achieved a discrimination accuracy of 80.65% and an F1-score of 79.42%. These findings can provide empirical support for real-time HAT assessment based on physiological features, contributing to the mitigation of flight safety risks resulting from inappropriate trust.

Data availability statement. The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy concerns.

Acknowledgements. The authors gratefully acknowledge the agencies NSFC, ZPNSFC and RFRPTL for the financial support. In addition, the authors acknowledge the participants for their participation.

Financial support. This research was funded by the National Natural Science Foundation of China (No. 52402507), Zhejiang Provincial Natural Science Foundation of China (No. ZCLMS26E0503), National Natural Science Foundation of China (No. 52272382) and Reward Funds for Research Project of Tianmushan Laboratory (No.TK-2024-D-012).

Competing interests. The authors declare no conflicts of interest.

Author contributions. Conceptualisation, formal analysis, methodology, software, funding acquisition, validation, visualisation and writing – original draft preparation, CY.F.; formal analysis, investigation, methodology, resources, software, validation and writing – review and editing, S.L.; conceptualisation, data curation, funding acquisition, methodology, project administration, resources, supervision and writing – review and editing, XR.WY.; investigation, methodology, software and validation, CY.Q.; investigation, methodology, software and validation, YQ.D.; formal analysis, visualisation, investigation, YR.L.; project administration, funding acquisition, resources, supervision, YM.Z. All authors have read and agreed to the published version of the manuscript.

References

[1] Lee, J.D. and See, K.A. Trust in automation: designing for appropriate reliance, *Human Factors*, 2004, **46**, (1), pp 50–80.
[2] Jamieson, G.A., Skraaning, G. and Joe, J. The B737 MAX 8 accidents as operational experiences with automation transparency, *IEEE Trans. Hum.-Mach. Syst.*, 2022, **52**, (4), pp 794–797. doi:10.1109/THMS.2022.3164774
[3] Saghafian, M., Vatn, D.M.K., Moltubakk, S.T., Bertheussen, L.E., Petermann, F.M., Johnsen, S.O. and Alsos, O.A. Understanding automation transparency and its adaptive design implications in safety–critical systems, *Saf. Sci.*, 2025, **184**, p 106730. doi:10.1016/j.ssci.2024.106730
[4] Walker, F., Forster, Y., Hergeth, S., Kraus, J., Payre, W., Wintersberger, P. and Martens, M. Trust in automated vehicles: constructs, psychological processes, and assessment, *Front. Psychol.*, 2023, **14**, p 1279271. doi:10.3389/fpsyg.2023.1279271
[5] Kohn, S.C., De Visser, E.J., Wiese, E., Lee, Y.-C. and Shaw, T.H. Measurement of trust in automation: a narrative review and reference guide, *Front. Psychol.*, 2021, **12**, p 604977. doi:10.3389/fpsyg.2021.604977
[6] Jian, J.-Y., Bisantz, A.M. and Drury, C.G. Foundations for an empirically determined scale of trust in automated systems, *Int. J. Cogn. Ergon.*, 2000, **4**, (1), pp 53–71.
[7] Campagna, G. and Rehm, M. A systematic review of trust assessments in human–robot interaction, *ACM Trans. Hum. -Robot Interact.*, 2025, **14**, (2), pp 1–35. doi:10.1145/3706123
[8] Zhang, T., Yang, J., Chen, M., Li, Z., Zang, J. and Qu, X. EEG-based assessment of driver trust in automated vehicles, *Expert Syst. Appl.*, 2024, **246**, p 123196. doi:10.1016/j.eswa.2024.123196
[9] Wang, M., Hussein, A., Rojas, R.F., Shafi, K. and Abbass, H.A. EEG-based neural correlates of trust in human-autonomy interaction, 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp 350–357. doi:10.1109/SSCI.2018.8628649
[10] Oh, S., Seong, Y., Yi, S. and Park, S. Neurological measurement of human trust in automation using electroencephalogram, *Int. J. Fuzzy Log. Intell. Syst.*, 2020, **20**, (4), pp 261–271. doi:10.5391/IJFIS.2020.20.4.261
[11] Oh, S., Seong, Y., Yi, S. and Park, S. Investigation of human trust by identifying stimulated brain regions using electroencephalogram, *ICT Expr.*, 2022, **8**, (3), pp 363–370. doi:10.1016/j.icte.2022.01.002
[12] Seet, M., Harvy, J., Bose, R., Dragomir, A., Bezerianos, A. and Thakor, N. Differential impact of autonomous vehicle malfunctions on human trust, *IEEE Trans. Intell. Transp. Syst.*, 2022, **23**, (1), pp 548–557. doi:10.1109/TITS.2020.3013278
[13] Chen, H. and Chan, I.Y.S. Effects of automation and transparency on human psychophysiological states and perceived system performance in construction safety automation: an electroencephalography experiment, *J. Constr. Eng. Manage.*, 2024, **150**, (7), p 04024055. doi:10.1061/JCEMD4.COENG-14205
[14] Blais, C., Ellis, D.M., Wingert, K.M., Cohen, A.B. and Brewer, G.A. Alpha suppression over parietal electrode sites predicts decisions to trust, *Social Neurosci.*, 2019, **14**, (2), pp 226–235. doi:10.1080/17470919.2018.1433717

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

[15] Perello-March, J.R., Burns, C.G., Woodman, R., Elliott, M.T. and Birrell, S.A. Using fNIRS to verify trust in highly automated driving, *IEEE Trans. Intell. Transp. Syst.*, 2023, **24**, (1), pp 739–751. doi:10.1109/TITS.2022.3211089

[16] Eloy, L., Doherty, E.J., Spencer, C.A., Bobko, P. and Hirshfield, L. Using fNIRS to identify transparency- and reliability-sensitive markers of trust across multiple timescales in collaborative human-human-agent triads, *Front. Neuroergonomics.*, 2022, **3**, p 838625. doi:10.3389/fnrgo.2022.838625

[17] Bandara, D., Liao, R., Chowdhury, F. and Abbott, L. Classification of human trust in AI using brain activity data, *Northeast J. Complex Syst.(NEJCS)*, 2025, **7**, (2), pp 1–27. doi:10.63562/2577-8439.1108

[18] Hopko, S.K. and Mehta, R.K. Trust in shared-space collaborative robots: shedding light on the human brain, *Hum. Factors*, 2024, **66**, (2), pp 490–509.

[19] Lee, J.R. and Gutzwiller, R.S. Do the eyes have it? A review of using eye tracking for automation trust measurement, *Human Factors*, 2025, **67**, (12), pp 1290–1311.

[20] Lu, Y. and Sarter, N. Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability, *IEEE Trans. Hum.-Mach. Syst.*, 2019, **49**, (6), pp 560–568. doi:10.1109/THMS.2019.2930980

[21] Zhang, Y., Yadav, A., Hopko, S.K. and Mehta, R.K. In gaze we trust: comparing eye tracking, self-report, and physiological indicators of dynamic trust during HRI, Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, 2024, pp 1188–1193.

[22] Hopko, S.K., Zhang, Y., Yadav, A., Pagilla, P.R. and Mehta, R.K. Brain–behavior relationships of trust in shared space human–robot collaboration, *ACM Trans. Hum. -Robot Interact.*, 2024, **13**, (1), pp 1–23. doi:10.1145/3632149

[23] Alzahrani, A., Nasir, J., Tayeb, A.J., André, E. and Ahmad, M.I. What do the face and voice reveal? Investigating trust dynamics during human-robot interaction, 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2025, pp 400–409.

[24] Ayoub, J., Avetisian, L., Yang, X.J. and Zhou, F. Real-time trust prediction in conditionally automated driving using physiological measures, *IEEE Trans. Intell. Transp. Syst.*, 2023, **24**, (12), pp 14642–14650.

[25] Hopko, S.K., Mehta, R.K. and Pagilla, P.R. Physiological and perceptual consequences of trust in collaborative robots: an empirical investigation of human and robot factors, *Appl. Ergon.*, 2023, **106**, 103863. doi:10.1016/j.apergo.2022.103863

[26] Kessler, T. *Neurophysiological Correlates of Trust in Robots*, University of Central Florida, 2020, Florida, FL: STARS.

[27] Ahmad, M. and Alzahrani, A. Crucial clues: investigating psychophysiological behaviors for measuring trust in human-robot interaction, International Conference On Multimodal Interaction, 2023, 135–143. doi:10.1145/3577190.3614148

[28] Yu, B., Bao, S., Zhang, Y., Sullivan, J. and Flannagan, M. Measurement and prediction of driver trust in automated vehicle technologies: an application of hand position transition probability matrix, *Transp. Res. Part C Emerg. Technol.*, 2021, **124**, p 102957. doi:10.1016/j.trc.2020.102957

[29] Yi, B., Cao, H., Song, X., Wang, J., Zhao, S., Guo, W. and Cao, D. How can the trust-change direction be measured and identified during takeover transitions in conditionally automated driving? Using physiological responses and takeover-related factors, *Hum. Factors: J. Hum. Factors Ergon. Soc.*, 2024, **66**, (4), pp 1276–1301. doi:10.1177/00187208221143855

[30] Yi, B., Cao, H., Song, X., Wang, J., Guo, W. and Huang, Z. Measurement and real-time recognition of driver trust in conditionally automated vehicles: using multimodal feature fusions network, *Transp. Res. Rec. J. Transp. Res. Board*, 2023, **2677**, (8), pp 311–330. doi:10.1177/03611981231156576

[31] Chang, W.-C., Gonzalez Garcia, N.F. and Hasanzadeh, S. Deep learning–based prediction of human–robot trust dynamics in future construction using worker neuropsychophysiological responses, *J. Comput. Civil Eng.*, 2025, **39**, (4), p 04025047. doi:10.1061/JCCEE5.CPENG-6302

[32] Campagna, G., Chrysostomou, D. and Rehm, M. Analysis of facial features for trust evaluation in industrial human-robot collaboration*, 2024 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO), 2024, pp 1–6. doi:10.1109/ARSO60199.2024.10557748

[33] Choo, S. and Nam, C. Detecting human trust calibration in automation: a convolutional neural network approach, *IEEE Trans. Hum.-Mach. Syst.*, 2022, **52**, (4), pp 774–783. doi:10.1109/THMS.2021.3137015

[34] Akash, K., Hu, W.-L., Jain, N. and Reid, T. A classification model for sensing human trust in machines using EEG and GSR, *ACM Trans. Interact. Intell. Syst.*, 2018, **8**, (4), pp 1–20. doi:10.1145/3132743

[35] Lu, Y. and Sarter, N. Modeling and inferring human trust in automation based on real- time eye tracking data, *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 2020, **64**, (1), pp 344–348. doi:10.1177/1071181320641078

[36] Wang, K., Hou, W., Ma, H. and Hong, L. Eye-tracking characteristics: unveiling trust calibration states in automated supervisory control tasks, *Sensors*, 2024, **24**, (24), p 7946. doi:10.3390/s24247946

[37] Santiago-Espada, Y., Myer, R.R., Latorella, K.A. and Comstock, J.R. The Multi-Attribute Task Battery II (MATB-II) software for human performance and workload research: A user's guide (NASA/TM–2011-217164). National Aeronautics and Space Administration,Langley Research Center, 2011.

[38] Sato, T., Jackson, A. and Yamani, Y. Number of interrupting events influences response time in multitasking, but not trust in automation, *Int. J. Aerosp. Psychol.*, 2024, **34**, (4), pp 208–224.

[39] Mognon, A., Jovicich, J., Bruzzone, L. and Buiatti, M. ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features, *Psychophysiology*, 2011, **48**, (2), pp 229–240.

[40] Feng, C., Liu, S., Wanyan, X., Chen, H., Min, Y. and Ma, Y. EEG feature analysis related to situation awareness assessment and discrimination, *Aerospace*, 2022, **9**, (10), p 546. doi:10.3390/aerospace9100546

[41] Hoff, K.A. and Bashir, M. Trust in automation: integrating empirical evidence on factors that influence trust, *Hum. Factors*, 2015, **57**, (3), pp 407–434. doi:10.1177/0018720814547570

[42] Billings, D.R., Schaefer, K.E., Llorens, N. and Hancock, P. What is trust? Defining the construct across domains, *Poster Pres. Am. Psychol. Assoc. Conf., Div.*, 2012, **21**, pp 1–7.

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press

[43] Hussein, A., Elsawah, S. and Abbass, H.A. The reliability and transparency bases of trust in human-swarm interaction: principles and implications, *Ergonomics*, 2020, **63**, (9), pp 1116–1132. doi:10.1080/00140139.2020.1764112
[44] Hancock, P.A., Kessler, T.T., Kaplan, A.D., Brill, J.C. and Szalma, J.L. Evolving trust in robots: specification through sequential and comparative meta-analyses, *Human Factors*, 2021, **63**, (7), pp 1196–1229.
[45] Li, Y., Wu, B., Huang, Y. and Luan, S. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust, *Front Psychol*, 2024, **15**, p 1382693. doi:10.3389/fpsyg.2024.1382693
[46] Taylor, R.M. Situation Awareness Rating Technique (SART): The development of a tool for aircrew systems design. 3-1-3–17, 1990.
[47] Trapsilawati, F., Herliansyah, M.K., Nugraheni, A.S.A.N.S., Fatikasari, M.P. and Tissamodie, G. EEG-based analysis of air traffic conflict: investigating controllers' situation awareness, stress level and brain activity during conflict resolution, *J. Navig.*, 2020, **73**, (3), pp 678–696. doi:10.1017/S0373463319000882
[48] Wanyan, X., Zhuang, D., Lin, Y., Xiao, X. and Song, J.-W. Influence of mental workload on detecting information varieties revealed by mismatch negativity during flight simulation, *Int. J. Ind. Ergon.*, 2018, **64**, pp 1–7. doi:10.1016/j.ergon.2017.08.004
[49] Armstrong, R.A. When to use the Bonferroni correction, *Ophthalmic Physiol Opt.*, 2014, **34**, (5), pp 502–508. doi:10.1111/opo.12131
[50] Feng, C., Liu, S., Wanyan, X., Dang, Y., Wang, Z. and Qian, C. β-wave-based exploration of sensitive EEG features and classification of situation awareness, *Aeronaut. J.*, 2024, **128**, (1329), pp 2561–2576.
[51] Bonnefond, M. and Jensen, O. The role of alpha oscillations in resisting distraction, *Trends Cogn. Sci.*, 2025, **29**, (4), pp 368–379. doi:10.1016/j.tics.2024.11.004
[52] Risqiwati, D., Wibawa, A.D., Pane, E.S., Yuniarno, E.M., Islamiyah, W.R. and Purnomo, M.H. Effective relax acquisition: a novel approach to classify relaxed state in alpha band EEG-based transformation, *Brain Inform*, 2024, **11**, (1), p 12. doi:10.1186/s40708-024-00225-y
[53] Frohlich, J., Toker, D. and Monti, M.M. Consciousness among delta waves: a paradox? *Brain*, 2021, **144**, (8), pp 2257–2277.
[54] Krausman, A., Neubauer, C., Forster, D., Lakhmani, S., Baker, A.L., Fitzhugh, S.M., Gremillion, G., Wright, J.L., Metcalfe, J.S. and Schaefer, K.E. Trust measurement in human-autonomy teams: development of a conceptual toolkit, *ACM Trans. Hum. -Robot Interact.*, 2022, **11**, (3), pp 1–58. doi:10.1145/3530874
[55] Wang, J., Fang, W., Qiu, H. and Wang, Y. The impact of automation failure on unmanned aircraft system operators' performance, workload, and trust in automation, *Drones*, 2025, **9**, (3), p 165. doi:10.3390/drones9030165
[56] Hergeth, S., Lorenz, L., Vilimek, R. and Krems, J.F. (2016). Keep your scanners peeled: gaze behavior as a measure of automation trust during highly automated driving. *Human Factors*, **58**, (3), 509–519.
[57] Ajenaghughrure, I.B., Sousa, S.D.C. and Lamas, D. Measuring trust with psychophysiological signals: a systematic mapping study of approaches used, *Multimodal Technol. Interact.*, 2020, **4**, (3), p 63. doi:10.3390/mti4030063

Cite this article: Feng C., Liu S., Wanyan X., Qian C., Dang Y., Lai Y. and Zhou Y. (2026). Sensitive physiological feature analysis related to human–automation trust assessment and discrimination. *The Aeronautical Journal*, **130**, 2808–2830. https://doi.org/10.1017/aer.2026.10182

https://doi.org/10.1017/aer.2026.10182 Published online by Cambridge University Press