frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 20 August 2025
DOI 10.3389/fcomp.2025.1549399

OPEN ACCESS

EDITED BY
Andrej Košir,
University of Ljubljana, Slovenia

REVIEWED BY
Güliz Tokadl,
Collins Aerospace, United States
Nuno Moura Lopes,
NOVA University of Lisbon, Portugal

*CORRESPONDENCE
Erin E. Richardson
✉ erin.richardson@colorado.edu

RECEIVED 21 December 2024
ACCEPTED 30 July 2025
PUBLISHED 20 August 2025

CITATION
Richardson EE, Kintz JR, Buchner SL, Clark TK and Hayman AP (2025) Operator-agnostic and real-time usable psychophysiological models of trust, workload, and situation awareness. *Front. Comput. Sci.* 7:1549399. doi: 10.3389/fcomp.2025.1549399

Operator-agnostic and real-time usable psychophysiological models of trust, workload, and situation awareness

Erin E. Richardson*, Jacob R. Kintz, Savannah L. Buchner, Torin K. Clark and Allison P. Hayman

Bioastronautics Laboratory, Ann & H.J. Smead Department of Aerospace Engineering Sciences, University of Colorado Boulder, Boulder, CO, United States

Trust, mental workload, and situation awareness (TWSA) are cognitive states important to human performance and human-autonomy teaming. Individual and team performance may be improved if operators can maintain ideal levels of TWSA. Predictions of operator TWSA can inform adaptive autonomy and resource allocation in teams, helping achieve this goal. Current approaches of estimating TWSA, such as questionnaires or behavioral measures, are obtrusive, task-specific, or cannot be used in real-time. Psychophysiological modeling has the potential to overcome these limitations, but prior work is limited in operational feasibility. To help address this gap, we develop psychophysiological models that can be used in real time and that do not rely on operator-specific background information. We assess the impacts of these constraints on the models' performance. Participants ($n = 10$) performed a human-autonomy teaming task in which they monitored a simulated spacecraft habitat. Regression models using LASSO-based feature selection were fit with an emphasis on model stability and generalizability. We demonstrate functional model fit (Adjusted R^2: $T = 0.67$, $W = 0.60$, SA $= 0.85$). Furthermore, model performance extends to predictive ability, assessed via leave-one-participant-out cross validation (Q^2: $T = 0.58$, $W = 0.46$, SA $= 0.74$). This study evaluates model performance to help establish the viability of real-time, operator-agnostic models of TWSA.

KEYWORDS

psychophysiology, predictive modeling, human-autonomy teaming, human-agent teaming, cognitive state

1 Introduction

Human-autonomy teaming is difficult in operational environments—environments characterized by trained users, ambiguity, hazards, and degraded performance and safety resulting from improper action, such as spaceflight, aviation, medicine, and military operations. Team performance and safety may be improved by the ability to predict an operator's trust (T), mental workload (W), and situation awareness (SA), three separate constructs collectively referred to as TWSA.

TWSA are constructs useful for understanding and predicting human-system performance (Parasuraman et al., 2008). Lee and See define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004). Maintaining appropriate trust is vital for avoiding misuse of over-trusted systems and for avoiding disuse of under-trusted systems (Lee and See, 2004; Akash et al., 2020). Longo et al. define mental workload as "the degree

of activation of a finite pool of resources, limited in capacity, while cognitively processing a primary task over time," emphasizing its mediation by external factors and characteristics of the human operator (Longo et al., 2022). Managing mental workload is important for optimizing the performance and enhancing engagement of operators (Longo et al., 2022; Heard et al., 2020). Endsley defines situation awareness as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988). Situation awareness is a critical input to decision making and is crucial to mission success (Endsley, 1988). Furthermore, Parasuraman et al. listed mental workload, situation awareness, and complacency, or over-trusting, as human performance consequences for evaluating automation design (Parasuraman et al., 2000). Methods of measuring or modeling these constructs can inform human-autonomy teams and improve their safety and performance.

Predictive models of TWSA must work in real time to account for changes in the environment and teammates or to facilitate adaptive autonomy if they are to be useful in real-world operations. These real-time predictive models would provide additional operational utility if they are accurate across different users. Collecting background information on operators may be impractical in time-constrained situations, like in the quick mobilization of fighter jets, in air traffic control centers, or in spaceflight missions where astronaut time is limited. An experiment on the International Space Station used reaction time tests and sleep-wake data to predict astronaut neurobehavioral performance, but was not able to collect reaction time test and sleep-wake observations daily (Tu et al., 2022). Additionally, a study assessing the effectiveness of rest periods on improving pilot alertness noted that in-flight data collection of reaction time task and performance data was restricted by operational demands (Rosekind et al., 1994). Furthermore, use of operator-specific data may decrease a technology's acceptability, especially in contexts where an operator's performance affects their career (Buenaflor et al., 2013). For example, concerns with sharing their personal data with managers was found to be an obstacle in construction workers' acceptance and intention to adopt wearable devices in the workplace (Choi et al., 2017). Since identifiable information facilitates mapping of data to individuals, avoiding collection of operator-specific information may help to decrease privacy risk and improve tool acceptability. Real-time, operator-agnostic models of TWSA are of critical interest as we look to improve the performance of human-autonomy teams in complex, constrained operational environments.

Current methods of measuring or predicting TWSA are insufficient for operational environments. Subjective questionnaires are the most widely-accepted method of measuring operator TWSA. Administering questionnaires, however, requires interrupting the work being performed, making them infeasible for deployment in operational settings. Others have attempted to measure cognitive states by monitoring behavioral actions a person takes during a task and parameters of the task itself, such as the number of times a participant rejects the autonomous system's suggested course of action, the time participants spent on a secondary task, and the time to notice a system's warning (Kintz et al., 2023b,c). Since behavioral and task-based measures are specific to a given task, they lose applicability with any changes to the task or to protocols on how operators complete the task. Additionally, accuracy in behavioral-based models of TWSA was shown to rely on operator-specific background information (Kintz et al., 2023b), discounting their potential for use in some constrained environments. For these reasons, questionnaires and behavioral-based measures of TWSA fall short in operational environments.

Psychophysiological measures provide a way to address the aforementioned challenges associated with predicting TWSA in operational settings. Non-invasive physiological responses may be recorded from the brain, heart, lungs, eyes, muscles, and skin and used in predictive models (Hettinger et al., 2003). Examples of physiological measures include heart rate, heart rate variability, respiration rate, tonic and phasic skin conductance, pupil diameter, and blink rate measures. Physiological measures do not require interrupting an operator's work and do not explicitly rely on task-specific elements. As a result, predictive models of TWSA based solely on physiological measures may show robustness to moderate changes in tasks or protocol as well as utility across different tasks.

Others have attempted to model operator cognitive states using physiological data; however, there are three primary limitations to prior work that the community should address prior to implementation: 1) models use proxy measures of cognitive states as ground truth; 2) only one cognitive state is modeled, reducing ecological validity; and 3) cognitive states are coarsely classified, limiting the usefulness of the measures. First, several studies use behavioral measures and aspects of the task itself as proxy measures of the cognitive states for the target variable to build their models (Kostenko et al., 2021; Liang et al., 2007; Zhou et al., 2022). While they may be related, these measures (e.g., task load) are not equivalent to the operator's true underlying cognitive state (which also depends on operator strategy, experience, fatigue, motivation, etc.). Further, behavioral measures and task elements are themselves readily observable and do not require prediction using psychophysiological signals. Second, studies tend to focus on one of TWSA rather than all three. This may be problematic since the constructs are interrelated and capable of changing simultaneously. Without independent measures of TWSA, it is difficult to discern associations between observable measures and each of TWSA. For example, reliance on autonomy can be indicative of high trust (Parasuraman et al., 2008), but it can also result from high workload, where participants turn to reliance due to a lack of mental resources (Biros et al., 2004; Kohn et al., 2021). This situation highlights why it is important to model the constructs independently to distinguish between ambiguous scenarios. Third, many efforts take a binary classification approach in modeling cognitive states (e.g., high vs. low workload), achieving less granularity in their predictions (Appel et al., 2019; Ferreira et al., 2014; Haapalainen et al., 2010; Liang et al., 2007; Zhao et al., 2018). Additionally, unlike regression modeling, machine learning models permit high-order interactions, requiring large amounts of data unless the signal to noise ratio is high (Harrell, 2017). We improve on prior work that relies on proxy measures by using validated questionnaires as the target variables in predicting TWSA. We also improve on work modeling singular cognitive

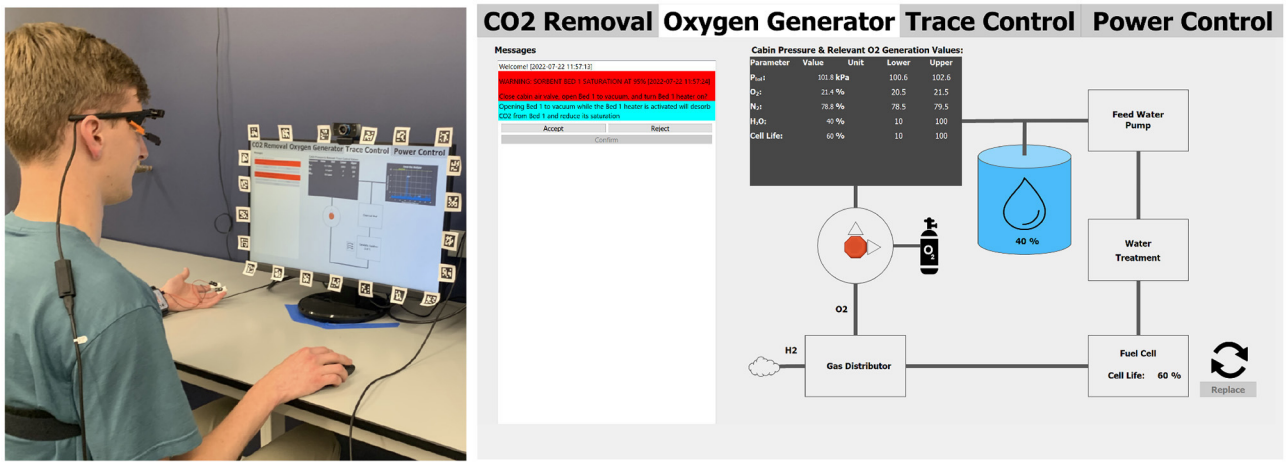

FIGURE 1
(a) Person wearing psychophysiological sensors (an electrocardiogram montage, a respiration belt, electrodermal activity electrodes, and eye-tracking glasses) while sitting at a computer. The person is engaging with a simulated space habitat maintenance task. **(b)** Picture of task screen on Oxygen Generator tab. On all tabs, the panel on the left side of the screen presents anomalies, suggestions from the simulated autonomous system, and options for actions the operator may take. The right side of the screen provides information on the states of the spacecraft subsystems (Kintz et al., 2023a).

states by predicting each of TWSA independently in one integrated task. Finally, our study's modeling approach improves upon classification by modeling TWSA as continuous constructs via multiple regression.

A key aspect of modeling human cognitive states using psychophysiology that has prohibited transition to operations is the necessity of comparing physiological signals to some baseline (e.g., an individual's average signal across the experiment), preventing them from being used in real time. Measuring changes from a physiological baseline can provide robustness to inter-individual differences, however, the best methods of normalizing signals likely differ between features. Further, there are a host of features that can be derived from physiological signals even beyond the means by which the data can be normalized or referenced to a baseline. The large number of candidate features combined with small sample sizes result in high-dimensional data. Dimensionality considerations are a primary challenge in psychophysiological modeling. Overfitting must be avoided to ensure models perform well on new data, but human subject experiments often have small sample sizes and either lack model validation or perform unreliable external validation on a small test set.

In this work, we explore the implications of excluding operator-specific information and features that cannot be used in real time on modeling TWSA in the same experiment. We build psychophysiological models of operator TWSA in a supervisory task where participants work alongside a simulated autonomous system. We evaluate the implications of normalizing physiological signals by extracting physiological features relative to different baselines to enable selection of the most useful features when ideal methods of baselining differ. We discuss our approach to building models from the resulting high-dimensional data. We perform feature shrinkage to reduce our feature set and stability selection to reduce variability in feature selection. Additionally, we use simulated random data to demonstrate our pipeline's robustness to false positives from high-dimensional data. Finally, we use internal cross-validation to assess the predictive accuracy of our models on unseen trials and on unseen people. This was done with different subsets of available features to assess the utility of real-time usable models and of operator-agnostic models. We hypothesized that model performance would decrease with less information available, but that real-time applicable, operator-agnostic models would still demonstrate viability for use in predicting operator TWSA.

2 Methods

In this experiment, participants worked in a supervisory role alongside a simulated autonomous system to maintain a modeled deep space habitat environmental control and life support system (ECLSS), as shown in Figure 1 and described in detail in Kintz et al. (2023a). Participants wore psychophysiological sensors during the experiment and self-reported their TWSA after each trial. We also collected background and demographic information from each participant. Predictive models of TWSA were built using participants' subjective questionnaire scores as the ground truth. This was done with different subsets of available features to assess the utility of real-time usable models and of operator-agnostic models. Each model's predictive accuracy was assessed through exhaustive cross-validation. An overview of our methods is presented in Figure 2.

2.1 Experiment design and protocol

This study was approved by the University of Colorado Boulder's Institutional Review Board (Protocol 21-0574).

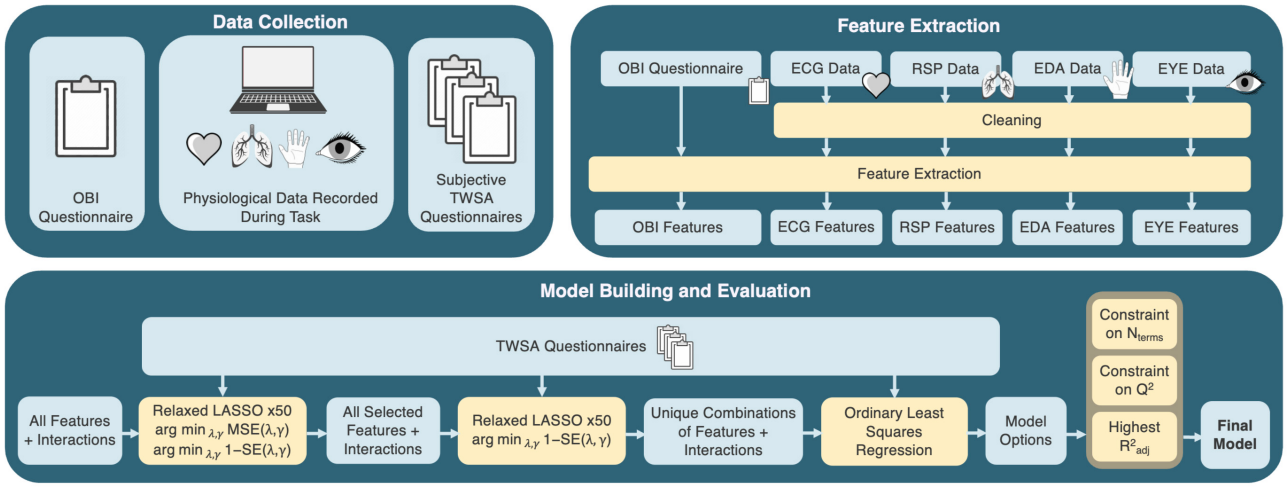

FIGURE 2
Overview of data collection, feature extraction, and model building and evaluation methods. Blue blocks represent data and results while yellow blocks represent processes. The data collection process is described in detail in Section 2.1, the feature extraction process is described in detail in Section 2.2, and the model building and evaluation process is described in detail in Section 2.3.

Participant exclusion criteria were: age outside of 18–65, consumption of alcohol in the 6 h prior to the study visit, known history of photosensitive epilepsy, known paraben allergy due to the use of electrode gels, and required use of eyeglasses to correct vision (contact lenses were acceptable) due to the use of eye-tracking glasses. Twelve people participated in the study but data from two of them were excluded due to technical issues during data collection. De-identified data from the remaining ten participants (6 male, 4 female, 19–42 years old, mean age of 24.8 years, standard deviation in age of 7 years) were used in our analysis. This sample achieved near gender parity, which is important as much of the literature, especially studies focused on pilots, have overwhelmingly male samples (Kostenko et al., 2021; Roscoe, 1984; Roscoe and Ellis, 1990; Wang et al., 2020; Zhou et al., 2022). Participants were paid for their participation and could earn a performance-based bonus in addition to compensation for their time. The experiment took place at the Human Research Laboratory at the University of Colorado Boulder's Aerospace Engineering Sciences building.

On a separate day from experiment data collection, each participant was trained on the task; the goal of the training was to ensure all participants reached proficiency prior to data collection to minimize learning effects during the experiment. All participants reviewed a slideshow of training content, completed a verbal quiz to ensure material comprehension, and completed at least 10 practice trials. Participants had the option to complete additional practice trials if they felt it would help improve their comprehension of the task. Before starting the experiment on the day of data collection, participants reviewed training materials, completed an operator background information (OBI) questionnaire concerning their quality of sleep the night prior, their dominant hand, and their experience with video games, robotic systems, aerospace displays, and environmental control systems. Participants also completed the Automation-Induced Complacency Potential (AICP) rating scale (Merritt et al., 2019). This measure aims to capture individual differences in inclinations toward complacency [suboptimal monitoring of an automated system that leads to performance failures (Parasuraman and Manzey, 2010)]. Next, participants completed the 3-min Psychomotor Vigilance Test (PVT), a measure of fatigue and alertness (Basner et al., 2015).

Next, participants donned psychophysiological monitoring equipment, pictured in Figure 3. Participants wore a 3-lead electrocardiogram (ECG) montage, a BIOPAC respiratory (RSP) chest belt, electrodermal activity (EDA) electrodes on two fingers of the hand that they do not use the computer mouse with, and Pupil Lab's Pupil Core (Berlin, Germany) eye-tracking headset (Kassner et al., 2014). The BIOPAC MP160 (Goleta, USA) system and BioNomadix loggers recorded ECG, RSP, and EDA signals at 2000 Hz. Cameras on the eye-tracking headset recorded each pupil at 120 Hz and recorded the participant's field of view at 30 Hz. These data streams, alongside trial start and stop times, were synchronized in Lab Streaming Layer. To provide baseline physiological reference values, the signals were recorded for 20 s before each trial while participants sat still and visually fixated on a crosshair on the computer monitor.

Participants worked alongside a simulated autonomous system to maintain a deep space habitat ECLSS (Kintz et al., 2023a). Throughout the trials, scripted anomalies occurred in the ECLSS system. The autonomous system recommended responses to each anomaly. These suggestions were incorrect in four randomly selected trials for each participant; faulty suggestions were implemented to evoke changes in participants' trust in the system across trials. The autonomous system had 5 modes of operation which varied system transparency and decision-making authority, aiming to elicit a range of cognitive states. The modes are shown in Figure 4a. Participants completed 15 trials, 3 with each autonomous system; the order in which they saw the modes was randomized. The trials were 50–95 s long, ensuring trials were long enough to assess a range of engagement levels while keeping the experiment length reasonable (~2 h) to manage participant

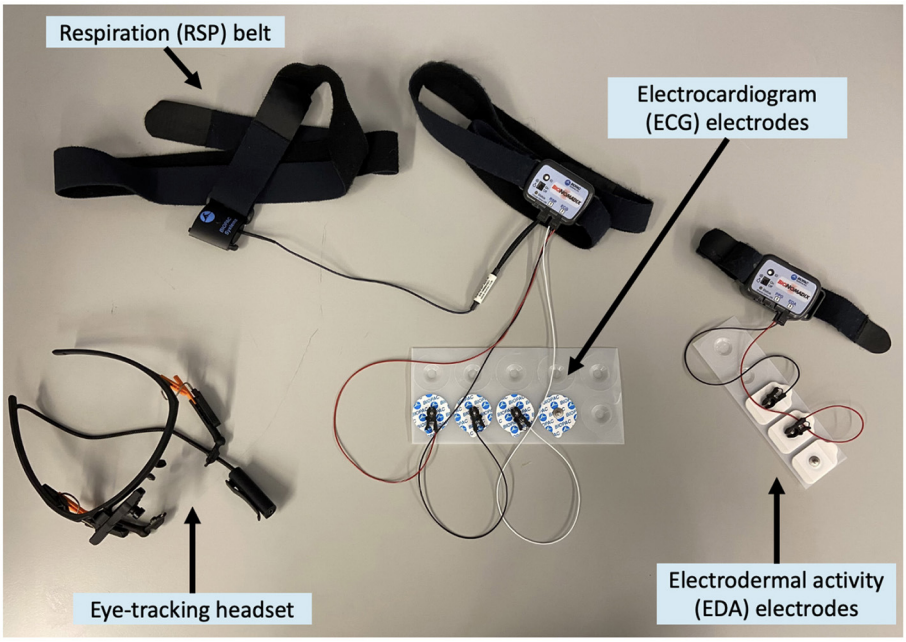

FIGURE 3
Psychophysiological sensors used in the experiment.

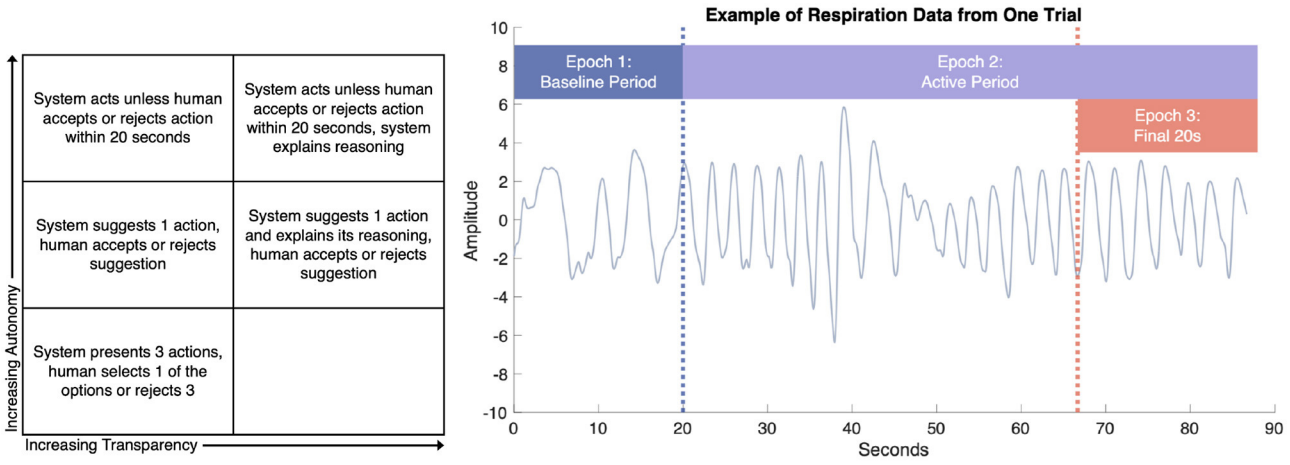

FIGURE 4
(a) Diagram of the autonomous system's five operational models. The modes increase in level of autonomy as you move up the diagram and increase in level of transparency as you move right across the diagram. Taken with permission from Kintz et al. (2023a). **(b)** A diagram indicating experiment trials and the epochs defined within trials. The diagram is overlaid with example respiration data collected during the experiment. The diagram outlines three epochs: a baseline period of 20 s before a trial starts, an active period during the trial, and the final 20s of a trial. These epochs were used for psychophysiological feature extraction.

questionnaire fatigue and storage of eye tracking recordings. Participants were paid for their time participating in the study and could also earn a performance-based bonus. Participants earned a small monetary bonus for every second that cabin air variables were within nominal bounds and would lose some of their bonus for every second that cabin air variables were out of nominal bounds. This reward system was implemented to increase stakes in the task and elicit vulnerability when relying on the autonomous system, providing a more realistic human-autonomy teaming scenario (Montague and Webber, 1965; Loft et al., 2023).

Participants self-reported their TWSA via subjective questionnaires after each trial. Jian et al.'s scale was used to measure trust (Jian et al., 2000), a modified version of the Bedford Workload scale was used to measure mental workload (Roscoe, 1984; Roscoe and Ellis, 1990), and the "10D" Situational Awareness

Rating Technique (SART) was used to measure situation awareness (Taylor, 1990).

One trial was excluded due to a participant not completing the modified Bedford scale and SART questionnaire. A total of $N_{\text{observations}} = 149$ were retained. In three instances, participants answered "10" on the modified Bedford workload scale after a trial where the autonomous system was unreliable. These trials were removed since participants' responses indicated that they could not complete the task due to the autonomous system, rather than due to extremely high workload. A total of $N_{\text{observations}} = 146$ were retained for use in the workload models while $N_{\text{observations}} = 149$ were retained for the trust and situation awareness models.

2.2 Feature extraction

2.2.1 Signal cleaning and raw features

The following data cleaning procedures were applied to the ECG data. First, we filtered the data using a highpass filter with a passband frequency of 1 Hz to remove baseline drift, a lowpass filter with a passband frequency of 100 Hz to remove electromyographic noise, and an infinite impulse response (IIR) Butterworth bandstop filter with a lower cutoff frequency of 59 Hz and an upper cutoff frequency 61 Hz to remove powerline interference (Jeyarani and Jaya Singh, 2010; Kher, 2019). Zero-phase digital filtering was used to eliminate the non-linear phase distortion of IIR filtering. R-DECO was used to extract R-peaks from the cleaned ECG signal, all of which were subsequently visually inspected and confirmed (Moeyersons et al., 2019). The following ECG features were extracted: heart rate, standard deviation of NN intervals (SDNN), percentage of NN intervals over 50 ms (pNN50), and root mean square of successive differences between heartbeats (RMSSD).

The following data cleaning procedures were applied to the respiration data. First, our team filtered the data using an IIR Butterworth bandstop filter with a lower cutoff frequency of 0.05 Hz and an upper cutoff frequency of 3 Hz to remove baseline drift and high frequency noise while preserving breathing rates between 3 and 180 breaths per minute. This replicates the Neurokit2 toolbox's implementation of Khodadad et al.'s (2018) work (Khodadad et al., 2018; Makowski et al., 2021). Zero-phase digital filtering was used to eliminate the non-linear phase distortion of IIR filtering. Then, the following RSP features were extracted: respiratory rate, median breath amplitude, the square of the change in chest circumference over a breath (a proxy for tidal volume), and the proxy for tidal volume multiplied by respiratory rate (a proxy for minute ventilation).

The following data cleaning procedures were applied to the EDA data. To identify motion artifacts, the EDA data was searched for values outside of 1 to 40 μS and for regions with high first or second derivatives (Braithwaite et al., 2013; Fowles et al., 1981). No regions were flagged for removal. Next, a Savitzky-Golay finite impulse response smoothing filter of polynomial order 3 was used to smooth the EDA signal (Savitzky and Golay, 1964; Thammasan et al., 2020). Ledalab's continuous decomposition analysis was used to decompose the EDA signal into tonic and phasic components (Benedek and Kaernbach, 2010). The EDA signal was first downsampled to 10 Hz, as recommended in Ledalab's documentation. The toolkit returned the deconvolved components and skin conductance response (SCR) characteristics. The following EDA features were extracted: mean, minimum, and maximum tonic skin conductance level, mean, minimum, and maximum phasic skin conductance level, number of SCRs per second, summed SCR amplitudes per second, and the area under SCRs per second.

Blink rate, blink duration, and average pupil diameter were measured by Pupil Labs' Pupil Core headset (Kassner et al., 2014) and retained as features. The responses to the OBI questionnaire, Monitoring subscores from the AICP scale, and PVT scores discussed in Section 2.1 were also provided as potential features to models that included operator-specific information.

2.2.2 Feature versions and interactions

Given the breadth of psychophysiological features available, it is likely that ideal methods of baselining differ between measures. Thus, several "versions" of each feature were calculated and provided as potential predictors.

Three epochs, or windows of time of interest, were defined for each trial, as shown in Figure 4b: 1) a 20 s baseline period before each trial where participants sat still and focused on a crosshair; 2) the entire 50–95 s active period where participants engaged with the ECLSS and autonomous system during a trial; and 3) the last 20 s of the active period immediately prior to trial end, as it represents the physiologic response immediately prior to filling out the TWSA questionnaires. Raw features, as described previously for each physiological signal, were calculated within these three epochs. These values were then used to create various "versions" of each feature. For example, one version subtracts the baseline period value from the active period value. Another version normalizes all values for a feature within the final 20 s epoch. In total, 23 versions of each feature were calculated. The versions are defined in the Supplementary materials. First-order interaction terms were also generated, further expanding the potential predictor space.

2.2.3 Feature subsets

The inclusion of different predictors drives the operational utility of a model. Buchner assessed the value of physiological signals alongside their burden to operators (Buchner et al., 2025). The total set of potential predictors defined in the present study can instead be categorized by their real-time applicability and their operator specificity.

We first divided the potential predictors into two subsets: batch applicable and real-time applicable. Batch applicable measures can be used after all data collection is complete and thus may use averages of all of a participant's trials or of all participants' data. Real-time applicable measures can be used as data is recorded and thus cannot rely on data that had not been recorded yet. We then divided the potential predictors into two more subsets: operator-specific and operator-agnostic. Operator-specific predictors include the OBI questionnaire responses, AICP scores, and PVT scores while the operator-agnostic features exclude them. Together, these subsets provided four model types and corresponding subsets of potential predictors.

2.3 Model building and evaluation

Ordinary least squares (OLS) multiple regression models of the form $y = \beta_0 + \beta_1 x_1 + ... \beta_p x_p + \varepsilon$ were fit for each model type for each of TWSA. We use the following process to account for the large set of potential predictors and relatively small number of observations to ensure an appropriate model building process, building from prior work (Buchner et al., 2025) and shown in Figure 2. All feature selection was done in R (Version 2022.12.0+353) while all data cleaning, feature extraction, and model building was done in MATLAB 2023a. First, sets of predictors $(x_1, x_2, ..., x_p)$ were down-selected from the subset of potential predictors for a given model type (e.g., real-time applicable and operator-specific). Using R's `glmnet` package, 10-fold cross validation relaxed LASSO was used to identify two sets of predictors by: 1) setting the shrinkage coefficient, λ, at the one standard error (1-SE) location and 2) setting λ at the minimum mean squared error (MSE) location (Meinshausen, 2007; Tibshirani, 1996). 10 folds were selected to balance bias and variance in the cross-validation (Breiman and Spector, 1992). This was repeated 50 times, creating 100 sets of predictors. Any terms that appeared in any of the 100 sets were used in a subsequent run of LASSO with λ at the 1-SE location. This was repeated 50 times, resulting in 50 more sets of predictors. This method of stability selection aims to reduce the instability in predictor sets resulting from cross-validation embedded in the LASSO method (Meinshausen and Bühlmann, 2010).

OLS was used to fit coefficients and a y-intercept to each unique set of predictors output by LASSO, generating a set of model options. Two forms of exhaustive cross-validation were performed to assess the model options. First, leave-one-participant-out (LOPO) cross validation performed the OLS step on nine participants and assessed predictive capability on the remaining one participant. This was done for each participant to generate a LOPO Q^2. Q^2 is analogous to R^2 but computed on the left out cross-validation data, such that it a measure of predictive, rather than descriptive, model performance. Second, leave-one-trial-out (LOTO) cross validation performed the OLS step on 148 trials and assessed predictive capability on the remaining one trial. This was done for each trial to generate a LOTO Q^2. It should be noted that in the cross-validation measures, the left-out participant/trial was not left out of the feature selection, such that the final model features remained the same while the fitted coefficients differed slightly with each left-out prediction. This limitation gives rise to the potential for internal validation bias. To assess our process' validity, we ran an ancillary study evaluating our model building pipeline's ability to generate well-performing models from randomized input data. A predictor matrix with the same dimensions as the largest predictor matrix used in the experiment (batch processing applicable and OBI predictors) was filled with randomized values. The relaxed LASSO pipeline selected the same model >100 times in a row. This model achieved an adjusted R^2 of 0.153, which is much lower than our real models achieved. All of the trials were used to fit the final coefficients and calculate the adjusted R^2 for each model option. To further protect against overfitting, two constraints were defined a priori to downselect model options:

1. $N_{\text{predictors}} \leq \frac{1}{5} N_{\text{observations}}$
2. $\min(\text{LOPO } Q^2, \text{LOTO } Q^2) \geq \text{Adj } R^2 - 0.2$

TABLE 1 Performance of trust models.

Model description		$N_{\text{predictors}}$	Adjusted R^2	LOPO Q^2	LOTO Q^2
Batch applicable	Operator-specific	29	0.71	0.54	0.63
	Operator-agnostic	25	0.67	0.58	0.61
Real-time applicable	Operator-specific	25	0.67	0.58	0.61
	Operator-agnostic	25	0.67	0.58	0.61

TABLE 2 Performance of mental workload models.

Model description		$N_{\text{predictors}}$	Adjusted R^2	LOPO Q^2	LOTO Q^2
Batch applicable	Operator-specific	24	0.66	0.48	0.56
	Operator-agnostic	23	0.65	0.45	0.54
Real-time applicable	Operator-specific	24	0.63	0.48	0.54
	Operator-agnostic	20	0.60	0.46	0.49

We constrained models to contain no more predictors than 1/5 of the number of observations to reduce model complexity and avoid overfitting. $N_{\text{predictors}}$ refers to the number of predictors in each model, where one predictor may be an interacted term. $N_{\text{predictors}}$ does not include a model's y-intercept. We also constrained models to have Q^2s within a reasonable range of the adjusted R^2, as disparity in these values is indicative of overfitting. The model option which achieved the highest adjusted R^2 while satisfying these constraints was selected as the final model. This process was repeated for the four model types (batch vs. real-time and operator-specific vs. operator-agnostic) for each of TWSA.

3 Results

Participants self-reported their trust, workload, and situation awareness via questionnaires. Trust scores can range from 12 to 84 (Jian et al., 2000), workload scores can range from 1 to 10 (Roscoe, 1984; Roscoe and Ellis, 1990), and situation awareness scores can range from −14 to 46 (Taylor, 1990). Since the developed models predict participant responses to these questionnaires, their predictions fall on the same scales. The trust, workload, and situation awareness model performances are summarized in Tables 1–3 respectively. Each table shows performance across batch applicable vs. real-time applicable models. Performance is further evaluated for operator-specific and operator-agnostic models within each category.

In Table 1, the same model is shown for the real-time applicable, operator-agnostic trust model as for the batch-applicable, operator-agnostic trust model and the real-time applicable, operator-specific

TABLE 3 Performance of situation awareness models.

Model description		$N_{predictors}$	Adjusted R^2	LOPO Q^2	LOTO Q^2
Batch applicable	Operator-specific	29	0.90	0.79	0.87
	Operator-agnostic	29	0.88	0.79	0.81
Real-time applicable	Operator-specific	29	0.85	0.74	0.78
	Operator-agnostic	29	0.85	0.74	0.78

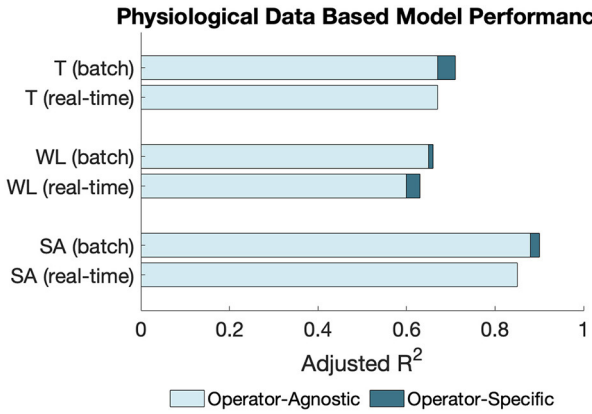

FIGURE 5
Comparison of model performances when limiting models to real-time applicable features and with inclusion and exclusion of OBI features. "T" indicates a trust model, "WL" indicates a workload model, and "SA" indicates a situation awareness model.

trust model. This model is more restrictive than the other two model types, and it outperformed those originally identified in the model selection process. As a result, the real-time applicable, operator-agnostic model was selected for all three of these model types. This model's adjusted R^2 was 0.67, only 0.04 lower than the batch-applicable, operator-specific trust model's adjusted R^2. Similarly, the predictive relevance of this model compared to the most inclusive model is improved by 0.04 points for LOPO Q^2 and 0.02 points lower for LOTO Q^2. Thus, these models of trust yield nearly equivalent performance.

Table 2 shows each of the 4 unique workload models identified. Workload model performance decreased as available potential predictors were removed, as is expected. Specifically, adjusted R^2 decreased by 0.06 between the least constrained (batch-applicable, operator-specific) and the most constrained (real-time applicable, operator-agnostic) workload models. Despite this slight decrease, the real-time applicable, operator-agnostic workload model still achieved good fit ($R^2 = 0.60$) and predictive relevance. The LOPO Q^2 measures were relatively stable, while the LOTO Q^2 dropped for the last, most restricted model. OBI was slightly more important to workload models when they were limited to real-time applicable features.

Table 3 shows model performance for each of the situation awareness model types. The situation awareness models yield the best performance across the TWSA models. Situation awareness model performance decreased as available potential predictors were removed, as is expected. Specifically, adjusted R^2 decreased by 0.05 between the least constrained (batch-applicable, operator-specific) and the most constrained (real-time applicable, operator-agnostic) situation awareness models. Despite this decrease, the real-time applicable, operator-agnostic situation awareness model still achieved good fit and predictive relevance. Including OBI features (i.e., operator-specific models) did not improve performance of the real-time applicable situation awareness models.

Comparative model performance is shown in Figure 5.

As seen in Figure 5, the situation awareness models consistently achieved the highest performance, followed by the trust models. Within each of TWSA, two bars are used to indicate the difference in goodness of fit (as measured by adjusted R^2) for the batch applicable and the real-time applicable models. The top bar for each construct shows the model goodness of fit when all predictors were made available, and the bottom bar shows goodness of fit when only the real-time applicable features were included. The darker blue bars show the goodness of fit of each model when OBI predictors are made available. Overall, limiting TWSA models to real-time applicable features did not substantially decrease performance. Furthermore, these real-time applicable models still performed well when OBI was excluded. The LOPO and LOTO metrics demonstrate that the models capture broad participant performance as well as predictive ability within participants.

Critically, the real-time applicable, operator-agnostic TWSA models still performed well when predicting the trust of participants whose data were not used to fit their coefficients, as indicated by their LOPO Q^2s of 0.58, 0.46, and 0.74, respectively. Figure 6 shows each of the three constructs' model performance using LOPO cross-validation. The x-axis shows participant-reported values and the y-axis shows model-predicted values. Each left-out participant is plotted with a different symbol. Values along the unity line indicate perfect prediction. Importantly, no model shows substantial bias or inability to capture the response from a given individual. This result is important as we aim to build models that can generalize to new operators.

4 Discussion

The real-time applicable, operator-agnostic models of trust, workload, and situation awareness demonstrated predictive relevance. The models' basis in psychophysiological signals as opposed to task-specific information and operator-specific information make them well-suited to predicting TWSA in operational environments. A primary contribution of this research is in emphasizing robust mathematical approaches to reduce the potential for model overfitting.

This is the first assessment of the viability of real-time applicable, operator-agnostic physiological models of cognitive states. Though prior work develops predictive models of cognitive states, they emphasize model accuracy over operational utility. Specifically, physiological signals that cannot be processed in real time (e.g., require comparing to the mean of responses over all

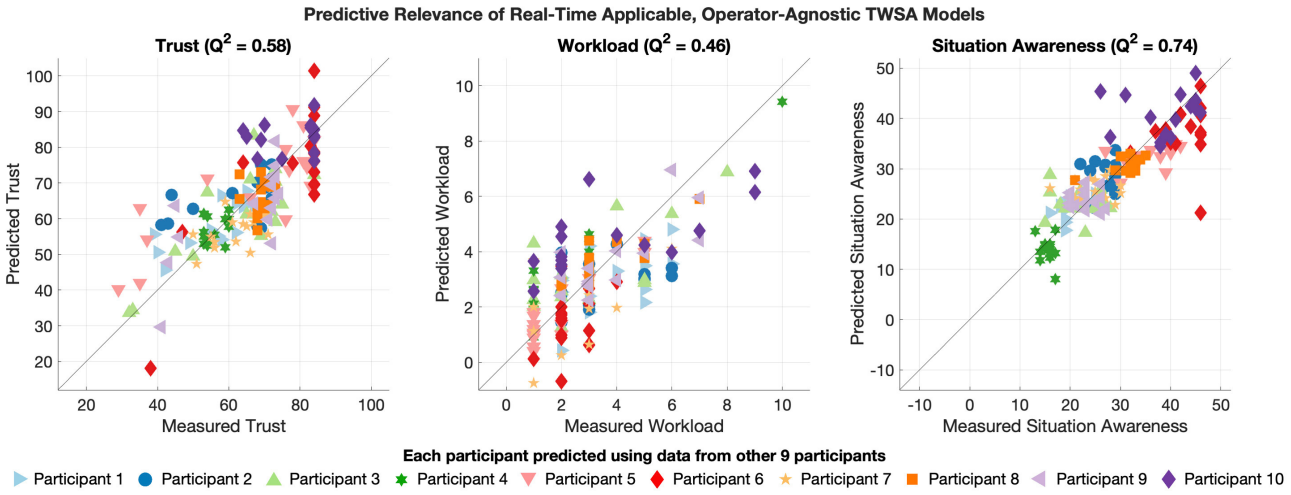

FIGURE 6
Leave-One-Participant-Out cross validation predictive performance for real-time applicable, operator-agnostic models of trust (left), workload (center), and situation awareness (right). Each construct is predicted on the scale of its subjective questionnaire, but the models may extrapolate beyond the scales' ranges.

trials) do not in fact offer a means for providing real-time estimates of cognitive states.

This study improves on prior work by modeling TWSA independently in one integrated task. Each of these validated constructs are important to human-autonomy teaming (Parasuraman et al., 2008), but are interrelated, making it important to model them separately. Importantly, this approach enabled a demonstration of the viability of the real-time applicable, operator-agnostic models for all three states.

The situation awareness models achieved the highest adjusted R^2 and Q^2 values across all model types. There may be several reasons why these were the best performing models. In particular, the data to which the model was trained was constrained to a portion of the total scale and was relatively stable within individuals. In this study, the operator was always acting in a supervisory manner. Kaber and Endsley investigated the influence of level of automation and adaptive automation on situation awareness and workload in a dynamic control task and found level of automation to be the driving factor in situation awareness (Kaber and Endsley, 2004). This study's results differ from the Kaber and Endsley study in that the operator was always acting in a supervisory manner. As such, operators consistently reported relatively high SA. Since the simulated autonomous system teammate varied its level of authority across trials, participants' roles in decision making, and thus, likely their understanding of the ECLSS system and how it was changing over time, varied across trials too, but perhaps not to the same extent seen in prior work.

The workload models achieved lower R^2 and Q^2 values than the trust and situation awareness models across all model types. In the same study, Kaber and Endsley found the proportion of time which their trials were automated to be the driving factor in perceived workload (Kaber and Endsley, 2004). Additionally, Röttger et al. found subjective ratings of operator workload to be significantly lower when operators were assisted with automation than in manual control conditions (Röttger et al., 2009). Notably, this study also saw operators maintaining a simulated spacecraft ECLSS in the presence of anomalies. In our study, participants dealt with anomalies, which were used to vary task load, but were always assisted by an autonomous system. Consistent with prior findings, participant reported values tended toward lower workload. Despite this, some participants did report high workload for some trials. Thus, we hypothesize the key factor reducing model performance is likely methodological. Unlike the questionnaires used to report trust and situation awareness, the modified Bedford scale is ordinal and is known to have uneven spacing between values on the scale (Casner and Gore, 2010). Our models predict workload on a continuous scale; this discrepancy likely contributed to the workload models' lower accuracy. This was done since the feature reduction methods (LASSO) operate on continuous, rather than ordinal data, which is a limitation to this approach. Future work should integrate ordinal regression within the stability selection LASSO pipeline implemented here to both model mental workload in better alignment with the construct and improve predictive accuracy.

Across all cognitive states, performance did not substantially decrease when OBI predictors were excluded. The largest decrease in adjusted R^2 was 0.04. Some decrease in model performance with the exclusion of predictors is expected, but these results suggest physiological signals may help retain performance. Previous work found trust and workload model performance to greatly decrease without OBI predictors (Kintz et al., 2023b). Critically however, these models did not have physiological measures available to them and only used task-embedded measures. In our study, only small drops in performance were seen with the exclusion of operator-specific predictors. For example, the batch-applicable trust models experienced a small drop in adjusted R^2 of 0.04 when operator-specific predictors were excluded. Lee and See explain that trust is influenced by systematic differences between people, and that individual propensities toward trust evolve alongside individual, organizational, and cultural context (Lee and See, 2004). While

the reduction in performance was minor, the literature helps explain the drop in trust model performance when OBI predictors are excluded.

Models using only real-time applicable predictors demonstrated nearly equivalent performance to models with batch-applicable predictors (the largest decrease in adjusted R^2 was 0.05). These models could be used for real-time prediction of TWSA to improve operator and team performance and safety. For example, adaptive autonomous systems can achieve greater performance by intelligently adapting their behavior based on a human operator's cognitive state (Feigh et al., 2012). In the case of a human operator and an autonomous system working together to supervise and maintain a space habitat, as simulated in this work, the autonomous system could decide to take on more of the task load when the operator's workload is too high, or it could decide to provide explanations for its suggested actions when the operator's trust is too low. Beyond adaptive autonomy, real-time prediction of TWSA could enable better management of human resources in workplaces, especially in operational environments where high performance is critical. A hospital team, for example, may be able to use predictions of individual workload to allocate personnel more efficiently between units (Momennasab et al., 2018). Air force squadrons could use real-time predictions of TWSA to provide pilots with dynamic understanding of their teammates' capacities, enabling better informed decision making. Both of these settings require fast responses from personnel, especially in the cases of medical emergencies or Air Force Quick Reaction Alerts, where the value in decision-making aids is limited by the time available for their use (Wohl, 1981). These scenarios further demonstrate the utility of real-time possible predictions of TWSA that do not rely on time-costly collection of OBI before use.

Beyond assessing model viability for operational environments, a key challenge addressed in this research is the emphasis on building approaches to deal with a large set of potential predictors and comparatively small sample size. These circumstances give rise to concerns regarding model stability and repeatability. Several efforts were made to validate both our model-building process and our selected models. Due to the interindividual variability in physiological measurements, normalizing via baselines was a major focus in model-building. Relaxed LASSO was chosen over LASSO in order to leverage its regularization and subsequent advantage in high-dimensional problems (Meinshausen, 2007). Second, stability selection, here implemented through the process of running relaxed LASSO on the set of predictors selected in prior runs of relaxed LASSO, was used to reduce falsely selected variables (Meinshausen and Bühlmann, 2010). Our ancillary study demonstrated our model building pipeline's inability to generate good models from randomized input data (with the same dimensionality as the real data).

There remain several areas of future work. Our work aims to generate useful insights into model applicability. As such, we are not attempting to develop models for deployment in operations, but rather provide an analysis to investigate the fundamental limitations of these kinds of models for operations. As such, prior to operational deployment, several additional steps must be taken. Despite our efforts to implement model-building and validation methods robust to small sample sizes, our small cohort is a limitation of our work (though consistent with other studies in the field). Additional data collection from larger sample sizes should be performed to ensure the most appropriate feature and the best models for predicting TWSA are identified. There are several limitations associated with subjective data use for model training. Participants' perceived cognitive states may not always be accurate. For example, participants may estimate their own situation awareness as high while not realizing that they are lacking understanding of some elements of the task. The self-report questionnaires are prone to subjectivity and may be confounded with other cognitive states (Selcon et al., 1991; Endsley et al., 1998). This experiment's lab-based setting and use of research-grade sensors limited the amount of noise in training data. The models should be tested in real-world operational settings to ensure their ecological validity. When transitioning beyond the lab, future work should evaluate the accuracy of these models with wearable-grade sensors. Wearable sensors would see increased noise but reduced pervasiveness, making them more feasible for use in field settings. Further, our population of participants drew primarily from university students. Real field operators would provide a more applicable sample to test models on before use. Note also that our cross-validation is limited in that the test data sets were not left out of the predictor selection process, giving rise to the potential for internal validation bias. In future work, this bias can be addressed by 1) collecting a larger dataset to enable data splitting and thus external validation on completely unseen samples and 2) including the feature selection step in the cross-validation to estimate the optimism in the model building process. Future work can address this by performing Monte-Carlo cross-validation to validate the entire model building process while building final models using all available data. It would also be valuable for future studies to assess which features are important for real-time, operator-agnostic predictions of operator cognitive states. Specifically, Shapley Additive Explanation (SHAP values), which quantify each feature's contribution to a particular prediction, could be aggregated across predictions to provide a measure of each feature's importance (Lundberg and Lee, 2017). This analysis could improve the interpretability of models and individual predictions, helping to increase transparency and trust in models proposed for use in high-stakes operational settings. Additionally, quantifying feature importance could inform the eventual down-selection of physiological sensors, which may be desirable to reduce sensor burden (Buchner et al., 2025).

In addition to externally validating these models on unseen trials or participants, it would be valuable for future work to assess the transferability of the models to new tasks. One major advantage of purely physiology-based predictions of TWSA is that they do not rely directly on measures specific to a given task, and thus insights on how to best predict TWSA (or even entire models) generated on one task may be useful in other tasks. The extent to which tasks can differ while TWSA models defined elsewhere remain useful should be investigated. Task variants (and their roles in modulating relationships between physiological measures and TWSA) of interest might include the nature of participants' actions (e.g., are they more physically, cognitively, or emotionally demanding?), the participant's level of authority, and the pressure the participant is under to perform well.

5 Conclusion

This work demonstrates the viability of real-time, operator-agnostic prediction of TWSA from psychophysiological data. During the experiment, operators maintained a simulated space habitat with an autonomous system teammate. We fit psychophysiological-based models to gold-standard questionnaire scores of TWSA, modeling three underlying cognitive states in one study. Performances of models subjected to different subsets of available predictors are compared, enabling informed trades of predictive accuracy and operational utility. TWSA models that were limited to measures that can be calculated in real-time and that were not privy to operator-specific OBI achieved good fit (adjusted R^2 of 0.67, 0.60, and 0.87 for TWSA, respectively) and predictive relevance (LOPO Q^2 of 0.58, 0.46, and 0.74 for TWSA, respectively). Additionally, the models are usable in real-time, and are unobtrusive in that they do not interrupt users' actions and do not rely on operationally-cumbersome OBI questionnaires. Furthermore, the psychophysiological predictors used in the models are task-agnostic, providing an enormous benefit of potential task transferability compared to task-specific measures used in other studies. Future work should assess the ability of these models to predict TWSA in different tasks and settings.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by University of Colorado Boulder's Institutional Review Board (Protocol 21-0574). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

ER: Conceptualization, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing, Formal analysis, Visualization. JK: Data curation, Investigation, Methodology, Software, Writing – review & editing. SB: Data curation, Investigation, Methodology, Software, Writing – review & editing. TC: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. AH: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by a Space Technology Research Institutes grant from NASA's Space Technology Research Grants Program under award number award number 80NSSC19K1052. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration (NASA).

Acknowledgments

An extended abstract presented at AAMAS 2024 describes a portion of this study (Richardson et al., 2024). We would like to thank everyone who participated in this study. We would also like to thank the Habitats Optimized for Missions of Exploration team for their collaboration and contributions. This work would not have been possible without the team of individuals who contributed to the development of the experiment task, especially Anna Jonsen, Maxwell Meiser, Ben Peterson, Josh Seedorf, Young-Young Shen, and Cody Wheeler.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2025.1549399/full#supplementary-material

References

Akash, K., McMahon, G., Reid, T., and Jain, N. (2020). Human trust-based feedback control: dynamically varying automation transparency to optimize human-machine interactions. *IEEE Control Syst. Mag.* 40, 98–116. doi: 10.1109/MCS.2020.3019151

Appel, T., Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., et al. (2019). "Predicting cognitive load in an emergency simulation based on behavioral and physiological measures," in *2019 International Conference on Multimodal Interaction, ICMI '19* (New York, NY, USA: Association for Computing Machinery), 154–163. doi: 10.1145/3340555.3353735

Basner, M., Savitt, A., Moore, T., Port, A., McGuire, S., Ecker, A., et al. (2015). Development and validation of the cognition test battery for spaceflight. *Aerosp. Med. Hum. Perform.* 86, 942–952. doi: 10.3357/AMHP.4343.2015

Benedek, M., and Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *J. Neurosci. Methods* 190, 80–91. doi: 10.1016/j.jneumeth.2010.04.028

Biros, D. P., Daly, M., and Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decis. Negot.* 13, 173–189. doi: 10.1023/B:GRUP.0000021840.85686.57

Braithwaite, J., Watson, D., Jones, R., and Rowe, M. A. (2013). *Guide for Analysing Electrodermal Activity and Skin Conductance Responses for Psychological Experiments*. Birmingham: CTIT Technical Reports Series.

Breiman, L., and Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *Int. Stat. Rev.* 60, 291–319. doi: 10.2307/1403680

Buchner, S. L., Kintz, J. R., Zhang, J. Y., Banerjee, N. T., Clark, T. K., and Hayman, A. P. (2025). Assessing physiological signal utility and sensor burden in estimating trust, situation awareness, and mental workload. *J. Cogn. Eng. Decis. Mak.* 19, 154–173. doi: 10.1177/15553434241310084

Buenaflor, C., Kim, H.-C., and Korea, S. (2013). *Six Human Factors to Acceptability of Wearable Computers*. International Journal of Multimedia and Ubiquitous Engineering. 8, 103–114.

Casner, S., and Gore, B. (2010). *Measuring and Evaluating Workload: A Primer*. Moffett Field, CA: NASA Ames Research Center.

Choi, B., Hwang, S., and Lee, S. (2017). What drives construction workers' acceptance of wearable technologies in the workplace?: indoor localization and wearable health devices for occupational safety and health. *Autom. Constr.* 84, 31–41. doi: 10.1016/j.autcon.2017.08.005

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proc. Hum. Factors Soc. Annu. Meet.* 32, 97–101. doi: 10.1177/154193128803200221

Endsley, M. R., Selcon, S. J., Hardiman, T. D., and Croft, D. G. (1998). A comparative analysis of sagat and sart for evaluations of situation awareness. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 42, 82–86. doi: 10.1177/154193129804200119

Feigh, K. M., Dorneich, M. C., and Hayes, C. C. (2012). Toward a characterization of adaptive systems: a framework for researchers and system designers. *Hum. Factors* 54, 1008–1024. doi: 10.1177/0018720812443983

Ferreira, E., Ferreira, D., Kim, S., Siirtola, P., Röning, J., Forlizzi, J. F., et al. (2014). "Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults," in *2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)* (Orlando, FL), 39–48. doi: 10.1109/CCMB.2014.7020692

Fowles, D. C., Christie, M. J., Edelberg, R., Grings, W. W., Lykken, D. T., David, T., et al. (1981). Publication recommendations for electrodermal measurements. *Psychophysiology* 18, 232–239. doi: 10.1111/j.1469-8986.1981.tb03024.x

Haapalainen, E., Kim, S., Forlizzi, J. F., and Dey, A. K. (2010). "Psycho-physiological measures for assessing cognitive load," in *Proceedings of the 12th ACM international conference on Ubiquitous computing, UbiComp '10* (New York, NY, USA: Association for Computing Machinery), 301–310. doi: 10.1145/1864349.1864395

Harrell, F. (2017). *Statistical Thinking - Classification vs. Prediction*. Statistical Thinking blog.

Heard, J., Fortune, J., and Adams, J. A. (2020). "SAHRTA: a supervisory-based adaptive human-robot teaming architecture," in *2020 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)* (Victoria, BC). doi: 10.1109/CogSIMA49017.2020.9215996

Hettinger, L. J., Branco, P., Encarnacao, L. M., and Bonato, P. (2003). Neuroadaptive technologies: applying neuroergonomics to the design of advanced interfaces. *Theor. Issues Ergon. Sci.* 4, 220–237. doi: 10.1080/1463922021000020918

Jeyarani, A. D., and Jaya Singh, T. (2010). "Analysis of noise reduction techniques on QRS ECG waveform - by applying different filters," in *Recent Advances in Space Technology Services and Climate Change 2010 (RSTS and CC-2010)* (Chennai), 149–152. doi: 10.1109/RSTSCC.2010.5712835

Jian, J.-Y., Bisantz, A., and Drury, C. (2000). Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* 4, 53–71. doi: 10.1207/S15327566IJCE0401_04

Kaber, D. B., and Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theor. Issues Ergon. Sci.* 5, 113–153. doi: 10.1080/1463922021000054335

Kassner, M., Patera, W., and Bulling, A. (2014). "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, UbiComp '14 Adjunct* (New York, NY, USA: Association for Computing Machinery), 1151–1160. doi: 10.1145/2638728.2641695

Kher, R. (2019). *Signal Processing Techniques for Removing Noise from ECG Signals*. jber.

Khodadad, D., Nordebo, S., Müller, B., Waldmann, A., Yerworth, R., Becher, T., et al. (2018). Optimized breath detection algorithm in electrical impedance tomography. *Physiol. Meas.* 39:94001. doi: 10.1088/1361-6579/aad7e6

Kintz, J., Shen, Y.-Y., Buchner, S., Anderson, A., and Clark, T. (2023a). "A simulated air revitalization task to investigate remote operator human-autonomy teaming with communication latency," in *52nd International Conference on Environmental Systems* Calgary.

Kintz, J. R., Banerjee, N. T., Zhang, J. Y., Anderson, A. P., and Clark, T. K. (2023b). Estimation of subjectively reported trust, mental workload, and situation awareness using unobtrusive measures. *Hum. Factors* 65, 1142–1160. doi: 10.1177/00187208221129371

Kintz, J. R., Buchner, S. L., Anderson, A. P., and Clark, T. K. (2023c). "Predicting operator cognitive states for supervisory human-autonomy teaming," in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* Oahu, HI: IEEE International Conference on Systems, Man, and Cybernetics (SMC). (in press). doi: 10.1109/SMC53992.2023.10394254

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., and Shaw, T. H. (2021). Measurement of trust in automation: a narrative review and reference guide. *Front. Psychol.* 12:604977. doi: 10.3389/fpsyg.2021.604977

Kostenko, A., Rauffet, P., and Coppin, G. (2021). Supervised classification of operator functional state based on physiological data: application to drones swarm piloting. *Front. Psychol.* 12:770000. doi: 10.3389/fpsyg.2021.770000

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors.* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Liang, Y., Reyes, M. L., and Lee, J. D. (2007). Real-time detection of driver cognitive distraction using support vector machines. *IEEE Trans. Intell. Transp. Syst.* 8, 340–350. doi: 10.1109/TITS.2007.895298

Loft, S., Bhaskara, A., Lock, B. A., Skinner, M., Brooks, J., Li, R., et al. (2023). The impact of transparency and decision risk on human–automation teaming outcomes. *Hum. Factors* 65, 846–861. doi: 10.1177/00187208211033445

Longo, L., Wickens, C. D., Hancock, G., and Hancock, P. A. (2022). Human mental workload: a survey and a novel inclusive definition. *Front. Psychol.* 13:883321. doi: 10.3389/fpsyg.2022.883321

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (Red Hook, NY, USA: Curran Associates Inc), 4768–4777.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., et al. (2021). NeuroKit2: a python toolbox for neurophysiological signal processing. *Behav. Res. Methods* 53, 1689–1696. doi: 10.3758/s13428-020-01516-y

Meinshausen, N. (2007). Relaxed lasso. *Comput. Stat. Data Anal.* 52, 374–393. doi: 10.1016/j.csda.2006.12.019

Meinshausen, N., and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B* 72, 417–473. doi: 10.1111/j.1467-9868.2010.00740.x

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., et al. (2019). Automation-induced complacency potential: development and validation of a new scale. *Front. Psychol.* 10:225. doi: 10.3389/fpsyg.2019.00225

Moeyersons, J., Amoni, M., Van Huffel, S., Willems, R., and Varon, C. (2019). R-DECO: an open-source Matlab based graphical user interface for the detection and correction of R-peaks. PeerJ. *Comput. Sci.* 5. doi: 10.7717/peerj-cs.226

Momennasab, M., Karimi, F., Dehghanrad, F., and Zarshenas, L. (2018). Evaluation of nursing workload and efficiency of staff allocation in a trauma intensive care unit. *Trauma Monthly* 23. doi: 10.5812/traumamon.58161

Montague, W. E., and Webber, C. E. (1965). Effects of knowledge of results and differential monetary reward on six uninterrupted hours of monitoring. *Hum. Factors* 7, 173–180. doi: 10.21236/AD0474470

Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Hum. Factors* 52, 381–410. doi: 10.1177/0018720810376055

Parasuraman, R., Sheridan, T., and Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* 30, 286–297. doi: 10.1109/3468.844354

Parasuraman, R., Sheridan, T., and Wickens, C. (2008). Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. *J. Cogn. Eng. Decis. Mak.* 2, 140–160. doi: 10.1518/155534308X284417

Richardson, E. E., Buchner, S. L., Kintz, J. R., Clark, T. K., and Anderson, A. P. (2024). "Psychophysiological models of cognitive states can be operator-agnostic," in *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, NZ), 2438–2440.

Roscoe, A. H. (1984). *Assessing Pilot Workload in Flight*. Technical report, Lisbon: Royal Aerospace Establishment Bedford.

Roscoe, A. H., and Ellis, G. A. (1990). *A Subjective Rating Scale for Assessing Pilot Workload in Flight: A decade of Practical Use*. Technical report, Royal Aerospace Establishment Farnborough: Farnborough.

Rosekind, M. R., Graeber, R. C., Dinges, D. F., Connell, L. J., Rountree, M. S., Spinweber, C. L., et al. (1994). *Crew Factors in Flight Operations 9: Effects of Planned Cockpit Rest on Crew Performance and Alertness in Long-haul Operations*. Moffett Field: Technical Report DOT/FAA/92/24.

Röttger, S., Bali, K., and Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. *Ergonomics* 52, 512–523. doi: 10.1080/00140130802379129

Savitzky, A., and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36, 1627–1639. doi: 10.1021/ac60214a047

Selcon, S. J., Taylor, R. M., and Koritsas, E. (1991). "Workload or situational awareness?: TLX vs. SART for aerospace systems design evaluation," *Proceedings of the Human Factors Society Annual Meeting* 35, 62–66. doi: 10.1518/107118191786755706

Taylor, R. M. (1990). "Situational awareness rating technique (SART): the development of a tool for aircrew systems design," in *Situational Awareness in Aerospace Operations (AGARD-CP-478)*. NATO - AGARD, Neuilly Sur Seine, France.

Thammasan, N., Stuldreher, I. V., Schreuders, E., Giletta, M., and Brouwer, A.-M. (2020). A Usability study of physiological measurement in school using wearable sensors. *Sensors* 20:5380. doi: 10.3390/s20185380

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tu, D., Basner, M., Smith, M. G., Williams, E. S., Ryder, V. E., Romoser, A. A., et al. (2022). Dynamic ensemble prediction of cognitive performance in spaceflight. *Sci. Rep.* 12:11032. doi: 10.1038/s41598-022-14456-8

Wang, H., Jiang, N., Pan, T., Si, H., Li, Y., and Zou, W. (2020). Cognitive load identification of pilots based on physiological-psychological characteristics in complex environments. *J. Adv. Transp.* 2020, 16. doi: 10.1155/2020/5640784

Wohl, J. G. (1981). Force management decision requirements for air force tactical command and control. *IEEE Trans. Syst. Man Cybern.* 11, 618–639. doi: 10.1109/TSMC.1981.4308760

Zhao, G., Liu, Y.-J., and Shi, Y. (2018). Real-time assessment of the cross-task mental workload using physiological measures during anomaly detection. *IEEE Trans. Hum.-Mach. Syst.* 48, 149–160. doi: 10.1109/THMS.2018.2803025

Zhou, F., Yang, X. J., and de Winter, J. C. F. (2022). Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. *IEEE Trans. Intell. Transp. Syst.* 23, 2284–2295. doi: 10.1109/TITS.2021.3069776