Real-time Cognitive State Estimates Modeling Using Embedded Measures for Adaptive Human-Autonomy Teaming

by

Santiago Huertas

B.S., University of Colorado Boulder, 2024

A master's thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Aerospace Engineering Sciences
2025

Committee Members:
Torin Clark, Chair
Allison Hayman
Katya Arquilla

Abstract

Huertas, Santiago (M.S., Aerospace Engineering Sciences)

Real-time Cognitive State Estimates Modeling Using Embedded Measures for Adaptive Human-Autonomy Teaming

Thesis directed by Associate Professor Torin Clark

It has been proposed that intelligent autonomous systems that dynamically estimate operator cognitive states can adapt their behavior to better aid human operators and enhance team performance. This capability is crucial for astronaut support during deep space missions, where communication delays with ground control limit timely assistance. Previous research on cognitive state estimation has generally focused on single states, observable factors (e.g., task load), and the use of multiple sensors simultaneously.

While our laboratory demonstrated an adaptive autonomous system that changed modes based on real-time estimates of trust, workload, and situation awareness, it exhibited limited accuracy potentially due to being trained on data from a non-adaptive system. Subsequent research performed a study that compared the predictive performance of multiple models with various feature availability variations but did not account for the effects of mode changes. To address this, we developed multiple models trained on data from subjects interacting with both a non-adaptive and an adaptive autonomous system. First, we trained models on the group that interacted with the adaptive autonomous system, incorporating users' background information, real-time sensor data, human-system interactions, and

novel features related to adaptation frequency and consistency of the autonomous system mode. Next, we trained and tested models built with exhaustive search followed by stepwise regression under optimizing for the Bayesian Information Criterion (BIC) using datasets from the adaptive group (N=10), the non-adaptive group (N=14), and a combined dataset (N=24). Finally, we explored three modeling methods with different feature down-selection approaches. Our results indicate that the combined dataset model focusing solely on human-system interaction features trained with exhaustive search followed by stepwise regression under BIC provided the most robust and best-performing prediction of operator trust, workload, and situation awareness. For example, the mean MAE (Mean Absolute Error) across 100 Monte Carlo Cross Validations for trust was 8.75, which was much smaller than the mean MAE by always predicting the median of the scale (as a baseline) which was 18.19. Similar results were found for models predicting workload and situation awareness.

This work enhances the state of the art in terms of the potential of adding new features that capture the interaction with an adaptive autonomous system, necessary datasets for training to have a more robust model when applied to unseen subjects, and approaches for building models to predict operator cognitive states for use in adaptive autonomous systems.

Acknowledgments

I am very grateful to my advisor, Dr. Torin Clark, for this opportunity, support, guidance, and patience along the way. I am fortunate to have worked with an advisor who is always available, supportive when things don't go well, helps me find a path when mistakes are made, is very passionate and transmits that passion to me, and encourages me to learn more and explore new paths. Thank you also to my mentor, Dr. Jacob Kintz, for being my guide and support in the research community. I am very grateful for everything I have learned and the opportunities I received from my advisor and mentor.

Thank you to Dr. Allison Hayman, for always having an open door, supporting me through my graduate school journey, and helping me navigate the research path.

Thank you to the many Bioastronautics students who provided insights and advice, as well as Dr. Katya Arquilla for serving on my thesis committee. Finally, thank you to my friends and family.

CONTENTS

Tables

Figures

Chapter 1: Introduction

Background and Motivation

In space missions, the challenge of maintaining optimal cognitive states is exacerbated by communication delays which are expected to last more than 20 minutes on deep space missions, and by constraints on available crew resources (Diamond et al., 2025; Calhoun et al., 2021; Parisi et al., 2023). These factors make real-time assistance from Mission Control unfeasible and increase the demand for autonomous systems that can mitigate the lack of knowledge or resource shortfalls (Anderson et al., 2020; Frank et al., 2016; Wischnewski et al., 2023). Human-autonomous system teaming becomes crucial as the crew needs to work together with the system to solve any abnormality or problem with their limited resources and abilities (Rollock & Klaus, 2022). By constantly monitoring and adjusting the cognitive states of an operator, such systems can modify their behavior to maintain high performance even under conditions of uncertainty (Borghetti et al., 2017; Byeon et al., 2025; Hancock et al, 2013).

Maintaining operators at their optimal levels of cognition is necessary to avoid issues like over-trust in automation, distrust, cognitive underload, overload, or loss of situation awareness all of which can compromise mission success (Dzindolet et al., 2002; Lee et al., 2004). For example, calibrated trust avoids both disuse and misuse of autonomous support so that the human operator will remain active and effective in the task (de Visser et al., 2018; Yang et al., 2021). Adaptive systems that change

their behavior dynamically as a function of unobtrusively estimated operator cognitive states will most likely improve multitasking and decision-making performance and reduce human error. Adaptive autonomous systems that have the capacity to adjust their mode or degree of transparency in real-time as a function of operator workload (W), situation awareness (SA), and trust (T) estimates and provide great benefits in novel and risky environments where crew performance is essential to ensure crew safety (Anderson et al, 2020).

To test the hypothesis that an adaptive autonomous system that can modify its behavior according to real-time cognitive state estimates will enhance human–system teaming, we created a computer-interface task that models an atmospheric revitalization system for a space habitat (Kintz, 2024). The task entails a simulated autonomous system that aids the operator in diagnosing and fixing any failure. Previous research has demonstrated that the incorporation of adaptive features into system design has the potential to drastically enhance operator performance and levels of safety in high-stakes domains (Heard & Adams, 2019). Furthermore, recent research has demonstrated that unobtrusive assessments based on physiological signals like ECG, electrodermal activity, and eye-tracking, as well as embedded behavioral assessments, can successfully quantify cognitive states in real-time (Schwarz &Fuchs, 2018 & Zhang, 2022; Kintz et al., 2023; Harrivel et al., 2017; Pereira et al., 2025). These methods offer a robust alternative to subjective questionnaires, which, although the gold standard for assessing cognitive states are

necessarily intrusive and not appropriate for ongoing measurement during task performance.

Cognitive states like trust, situation awareness, and workload are complex constructs that play a vital role in affecting decision-making and performance. Trust is a cognitive state which Lee and See defined as "…the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004). Situation awareness, as perception, comprehension, and projection of the features of the environment, is crucial to ensuring operational safety. Workload is the equilibrium between task demands and mental resources available; underload and overload are both performance disruptive and, at the extremes could lead to mission failure (Hancock & Matthews, 2018; Heard & Adams, 2019). Mismatches in these cognitive states, caused by excess workload, poor situation awareness, or miscalibrated trust, have been shown to cause error and decreased mission effectiveness (Parasuraman et al., 2008; Stanton et al., 2001). The incorporation of adaptive autonomous systems in demanding environments, including deep space missions, demands the development of robust, real-time models of operator cognitive states. By exploring embedded measures recorded in the experiment developed by Kintz et al., 2023b and Kintz, 2024, this study seeks to create integrated models that concurrently predict workload, trust, and situation awareness. These models not only improve our knowledge of the dynamics between various cognitive states but also open the possibility of creating autonomous systems that can dynamically adapt their behavior. For instance, when an operator's workload

is estimated to be approaching a critical level, the system can modify its transparency level or change its intervention mode in order to reduce cognitive load (Heard et al., 2020; Luo et al., 2021). This adaptive behavior is essential in deep space travel, where the environment is unknown, and dynamic, with limited resources and the penalty for mistakes is extremely high (Frasheri et al., 2018).

Finally, this thesis aims to contribute to the area of human–autonomy teaming through the development and verification of adaptive models that leverage unobtrusive measures to predict cognitive states in real-time by combining expertise in research-embedded performance measures, and adaptive control, we seek to create systems that not only assist operators in performing tasks in unknown environments but also optimize overall mission performance. The results of this research will contribute to the general body of knowledge regarding human-autonomous machine collaboration and shape the design of autonomous systems that are adaptive, resilient, robust, and able to keep cognitive states in optimal ranges despite adverse operating conditions.

Previous Work

Recent work suggests that adaptive autonomous systems that adjust their behavior in real-time based on operator cognitive states estimated using unobtrusive measures can considerably increase user performance and safety in mission-critical environments. For example, Anderson et al., 2020 suggested the potential of adaptive autonomy for future spacecraft habitats by proposing how dynamic system

adjustments can support operator decision-making under uncertainty, pressure, and lack of resources. Guo & Yang, 2021, propose a dynamic trust model capable of adjusting over time as the user interacts with the autonomous system using Bayesian inference. Feigh et al., 2012 provided a comprehensive framework for characterizing adaptive systems, emphasizing the importance of continuous monitoring of workload and situation awareness. De Visser et al., 2018 further explored the evolution from traditional automation to autonomy, demonstrating how dynamic trust repair mechanisms can be utilized to maintain trust in optimal ranges among operators. Additionally, Heard et al., 2020 introduced the SAHRTA (supervisory-based adaptive human-robot teaming architecture), which leverages multimodal sensor data to dynamically adjust the level of autonomy of the system or change the interface to allow for more meaningful interaction with the user to improve performance. Finally, Kintz et al., 2023 demonstrated that predictive models built on data from adaptive systems can more accurately estimate operator cognitive states while interacting with an autonomous system, thereby enabling effective real-time human-autonomy teaming.

Previous research on adaptive autonomous systems has considered mostly only a single cognitive state, adapted based on observable metrics (e.g., task load), and employed multimodal sensors simultaneously to estimate these states (Harrivel et al., 2017). Our previous experiments demonstrate that adaptive autonomous systems significantly impact users' cognitive states (Kintz, 2024). However, earlier models were trained using human subject data from experiments where subjects interacted

with a non-adaptive autonomous system, leading to limited accuracy when applied to subjects using an adaptive system (Kintz, 2024; Rote et al., 2024).

Prior research conducted by Rote et al., 2024 performed an ablation study to determine which feature combinations enhance accuracy in dynamic estimates of trust, workload, and situation awareness when predicting unseen participants. The study categorized the predicted features into five types: operator background information, agent mode, simulation events, operator actions, and operator eye tracking. It found that the best accuracy predictions resulted from including only agent mode, simulation events, and operator actions. Models for the three cognitive states were developed allowing features in just these categories, and the model will be referred to as "Model 6". However, because the training data came from a non-adaptive autonomous system, no features capturing the effects of interacting with an adaptive system were included. Furthermore, the data used to build the models does not exactly reflect what will be encountered when the model is used by an intelligent autonomous system, as subjects will experience changes in the system's mode both across and within trials.

There has not been a direct comparison of the performance of Model 6 using adaptive (Naive) versus non-adaptive (Static) training data, nor an examination of the effect of dataset size on model accuracy. Previous research developed models using exhaustive search followed by stepwise regression optimized by either BIC (Bayesian Information Criterion) (Kintz, 2024) or LASSO (Least Absolute Shrinkage

and Selection Operator) (Buchner et al., 2025; Richardson et al., 2024). However, there has not been a direct comparison between the performance of these methods and an alternative approach that employs exhaustive search followed by stepwise regression optimized by AIC (Akaike Information Criterion), which allows for the selection of more features.

Research Objectives

We tried to develop an improved model ("Best model") for cognitive states prediction on unseen subjects by looking at gaps in the current research, comparing modeling methods, and creating new unexplored features. Based on this, the goals of this research are as follows:

Aim 1: Develop new features that are focused on trying to capture the impact of interacting with an adaptive autonomous system. Compare the performance of models created that include these new features and interactions with models using just the prior features to see if we can improve model performance by reducing the mean absolute error of model predictions of cognitive states.

Aim 2: Compare the effects of data size and model performance when trained with Static (14 subjects) vs Naïve (10 subjects) vs Both (Static + Naïve = 24 subjects). Compare the model performance of a downselection of Both (N=24 to using only N=10) to a downselection of Static (N=14 to using only N=10) to Naïve (N=10) to have the same training dataset size, as well as with Both downselected to 14 to compare with Static.

Aim 3: Compare the model performance of multiple feature selection approaches. It includes an exhaustive search followed by stepwise regression optimized by either BIC (Bayesian information criterion) or AIC (Akaike information criterion) or using the model building LASSO (Least Absolute Shrinkage and Selection Operator).

Through the completion of these three Aims, we conclude with a single, final model that yielded the best and most robust model predictive accuracy, which we suggest should be used going forward for making real-time estimates of operator cognitive states when interacting with an adaptive autonomous system.

Chapter 2: Methods

Simulation Environment and Task

This thesis leverages previously collected human subject data capturing operator mental workload, trust, and situation awareness (Kintz et al., 2023; Kintz, 2024). The details of the experiment, the task that operators performed, and the data collected can be found in Kintz et al., 2023b, but are briefly summarized here to provide context. The supervisory human-autonomy teaming experiment was split into two sessions. First participants were trained on how an Environmental Control and Life Support System (ECLSS) worked and what are the corrective courses of action in case of any anomaly. The simulated environment included 4 tabs representing three support systems and an embedded secondary task: CO2 Removal, Oxygen Generator, Trace Control, and Power Control. They performed 10 training trials in which they experienced how the system would communicate with them to solve anomalies in an ECLSS system in a remote spacecraft with a 10 s latency. The experiment had CU-Boulder IRB approval.

The participant's task was to make the correct decision to solve any anomalies in any of these systems while working together with an autonomous system in a supervisory role. The user was only allowed to see the current states of the ECLSS but had no manual control. The autonomous system had different modes that changed the level of transparency or autonomy in how it provided information on the anomaly and the corrective action. Participants performed 15 trials, four of which the simulated autonomous system offered incorrect solutions to replicate a real-world

imperfect autonomous system and elicit realistic trust dynamics (Lee & See, 2004; Wickens & Dixon, 2007). They had an economic incentive if all parameters were in the nominal range and the more times they completed the secondary task. At the end of the trial, the simulation "freezes" and three questionaries pop up, which the participant had to fill out before seeing the results of their actions during the trial.

Prediction targets

Participants completed three surveys at the end of each trial to report their cognitive states corresponding to the final moment at the end of the trial (rather than over the entirety of the previous trial). The surveys correspond to the Trust in Automated Systems (TAS) for perceived trust, the Modified Bedford Scale for workload, and the Situation Awareness Rating Technique (SART) for situation awareness.

Table 2.0.1: Description of Prediction Targets

Metric	Description	Range
TAS (Jian et al., 2000),	Trust in Automated Systems – Ground truth for trust	12-84
Modified Bedford (Roscoe & Ellis, 1990),	Score from Modified Bedford Ordinal Scale - Ground truth for workload	1-10
SART (Selcon & Taylor, 1990).	Situation Awareness Rating Technique – Ground truth for situation awareness	-14-46

Predictor Variables

The predictor variables that may potentially be included in models of cognitive states can be split into three categories: agent mode, simulation events, and operator actions from "Model 6" by Rote et al. (2024). Table 2.2 shows all the predictors corresponding to each category, description, and their ranges.

Table 0.2.2: Description of Predictor Variables

Category	Variable	Value
Agent Mode	Level of Autonomy	Low: -1 Medium: 0 High: 1
Agent Mode	Level of Transparency	No explanation: 0 Explanation Given: 1
Simulation Events	Number of Events in a Trial	Min:1 Max:6
Simulation Events	Time Since Last Event	Min: 0.35s Max: 82.18s
Simulation Events	Trial Length	Min: 50.76s Max: 96.09s
Operator Actions	Time Since Most Recent Confirm	Min: 0.20s Max: 12.29s
Operator Actions	Number of Rejected Options	Min: 0 Max: 6
Operator Actions	Number of Tab Switches	Min: 2 Max: 31
Operator Actions	Fraction of Time Spent in Secondary Task	Min: 0 Max: 0.98
Operator Actions	Selected Accept and Reject in Same Trial	Yes: 1 No; 0
Operator Actions	Switched Tabs After Warning and Before Confirm Action	Yes: 1 No: 0
Operator Actions	Time Since Last Checked a Tab	Min: 0s Max: 46.21s

Table 2.2 includes the predictor variables selected by Rote, et al. (2024) for Model 6 which had the lowest Mean Absolute Error. These predictor variables were our starting point as we developed new features that considered the impact of interacting with an adaptive autonomous system (see the new features development section below).

Participants Groups

The data used in our analysis comes from two experiments that had participants interact with the same autonomous system, with the same 15 trials, simulation events, and consistent setup. The first dataset comes from Kintz et al, 2023 which had 14 participants. This group will be referred to as Static because participants were only presented with the autonomous system in a single mode across all warnings for each given trial. It means they only experienced 1 of the 5 possible system modes during each trial. However, this mode could change between trials. The second dataset comes from Kintz, 2024 which had 10 participants. This group will be referred to as Naïve because participants were presented with a random mode every time there was a warning as if the autonomous system was "naively" selecting a random mode each time. This serves as a useful dataset to capture any potential impact of the modes changing across warnings, even within a given trial (which would also occur in an adaptive autonomous system, which changes its mode based upon a real-time estimate of the operator's cognitive states).

Monte Carlo Cross Validations

The model performance, in terms of predictive accuracy, was quantified by comparing the Mean Absolute Error (MAE) when making predictions for "test" data in unseen subjects. Data sets were split using a 7/3 split, in which the data from 7 participants was used to train the model (both for feature selection and coefficient fitting) while 3 participants acted as the unseen test data. 100 Monte Carlo cross-validations (MCCVs), in which the train/test split was randomly performed for each

model, to capture most of the possible combinations of a 7/3 split. Across the 100 MCCVs, it is possible to capture the range and distribution of MAE outcomes, given the model accuracy will vary for each train/test split.

As done in Kintz et al. 2024, our first approach to model building was that models were trained by providing all possible features, then features were downselected by an exhaustive search that allowed to generate models without restrictions on the number of features but limited to not including any combinations between them (i.e., no interactions). Then the "best" model was selected by the combination of predictor main effects (i.e., allowing for interactions) with the lowest Bayesian Information Criteria (BIC). Then the "best" model was used to seed a stepwise regression optimized by the minimizing model's BIC. As an alternative, here we considered another approach that was identical, but optimized for the Akaike Information Criterion (AIC). Models were forced to include the Level of Autonomy and Level of Transparency in all considered possible models. The final feature set, interactions, and coefficients corresponded to the model with the lowest AIC/BIC.

New features Development

Previous research developed Model 6 training on the dataset which we term Static data. However, the actual intelligent adaptive system will be changing modes during and across trials. Figure 2.1 shows the performance of this mode when tested using Monte Carlo cross-validations trained using Naive data which is more representative of the data the model will be exposed to (i.e., if the predictions are being used to drive an adaptive autonomous system's mode change (Kintz, 2024), in

that the mode will be changing during trials). When compared with just taking the median of the scale for all cognitive states (which serves as a simple, completely data-agnostic prediction to which other prediction accuracies can be compared), Model 6 trained with Naïve data outperforms with a considerably lower mean MAE, as expected. When compared to the Model 6 trained with Static data, situation awareness (Figure 2.1c) is very similar with a slightly lower MAE for Model 6 with Naive, but with a similar distribution. Trust (Figure 2.1b) has a very similar mean MAE, but a wider range of the MCCV distribution when using Naive data in comparison with the Static data. A narrower range about the mean MAE represents a model that is more consistent in prediction accuracy for various train/test splits (i.e., which subjects are unseen in the model building).

Workload had the highest performance difference between using Static vs. Naive data. The mean MAE increased substantially, rising from 1.25 (Static) to 1.68 (Naive), with the distribution shifting toward the extremes of the range. In these comparisons (and the ones below), we intentionally did not perform statistical inference tests to determine if there was a significant difference in MAE performance between different models. This is somewhat of a flawed endeavor since significance can likely be achieved simply by increasing the number of MCCVs (producing a new dot in the violin plot simply through more computational effort) to increase the degrees of freedom until a negligible difference could reach significance. Instead, we focus on substantial, meaningful differences by visualizing the distribution of MAEs across MCCVs.

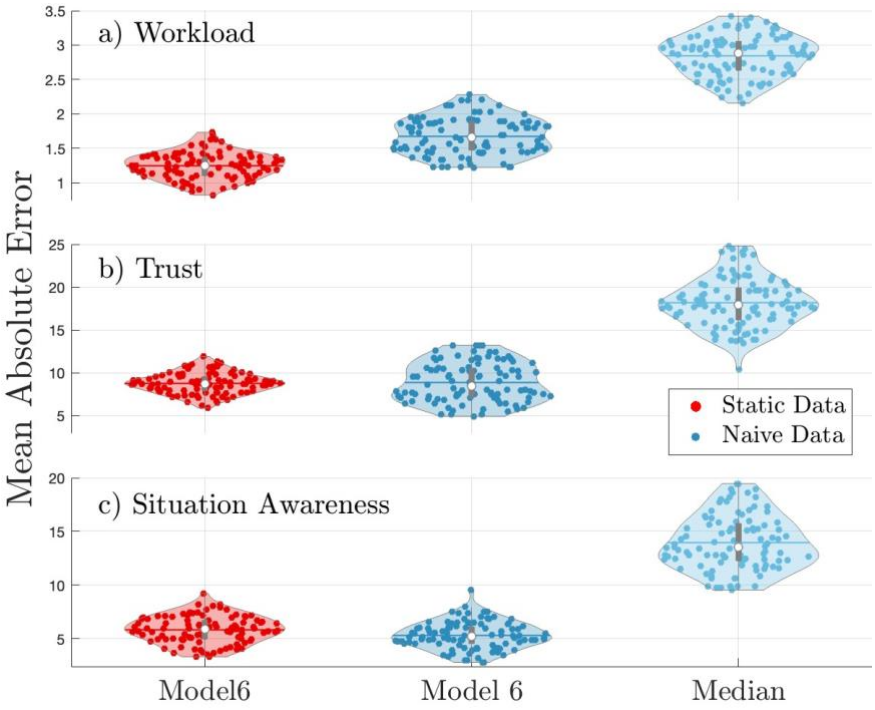

Figure 2.1 Workload (panel a), trust (panel b), and situation awareness (panel c) (or WTSA) predictive performance comparison using the model with the best performance (lowest mean absolute value) from the ablation study performed by Rote et al. 2024 (Model 6), with either data from subjects interacting with the non-adaptive system (Static, in red) and random adaptive system (Naïve, in blue) training data. Accuracy performance is compared to a baseline computed by always predicting the Median of the metric's scale for each cognitive state (i.e., a completely data-agnostic guess of the cognitive state). Each dot corresponds to the mean absolute value of predictive error from a Monte Carlo Cross Validation (MCCV) trained with 7 subjects and then applied to 3 unseen "test" subjects for Static and 10 and 3 subjects for Naïve, respectively. This distribution is visualized with a violin plot in shading, the mean of the 100 MCCVs with a horizontal-colored line, the white dot represents the median MAE across the 100 MCCVs, and the gray box represents the interquartile range (Bechtold, 2016).

The main difference between the Static and Naïve data sets is the changing autonomous system mode during and across trials. The performance shown in Figure 2.1 suggests that a Model 6 trained (and tested) using Naïve Data may be able to be

improved when the model builder can choose features that account for this (particularly for workload, where the model using Naive data had a higher mean MAE than that using Static data). Thus, we developed new features that tried to account for the effects of an adaptive system changing modes during trials (since that occurred in the Naive Data cohorts). Table 2.3 describes four new features and the scores. Conceptually, these considered short vs. longer timescales, either just whether the last two warnings were presented consistently in the same mode, how consistent modes were for warnings across the full trial or consistency in modes for warnings across the prior three trials.

Table 2.3: Description and Possible Values for New Features

Feature	Description	Value
Last 2	Compares the consistency of the autonomous system mode across the last 2 warnings in the same trial	Same Mode: -1 Only One Warning: 0 Different Modes: 0.25
Last 2.2	Compares the consistency of the autonomous system mode across the last 2 warnings even if they happened across different trials	Same Mode: -1 Only One Warning: 0 Different Modes: 0.25
Expected Modes (EM)	Compares the consistency of the autonomous system by comparing the expected number of modes if selected randomly to the experienced number of modes in the same trial	EMR(#W)-Experienced Modes Min: -1.44 Max: 1.05
EM.2	Compares the consistency of the autonomous system by summing the score from expected modes across the last three trials	EMR(#W)-Experienced Modes Min: -1.44 Max: 1.61

The expected number of modes based on the number of experienced warnings, if selected randomly, is defined by equation 2.1. As one example of how this feature works, if there were two warnings, the expected number of unique modes that occur is $5*(1-0.8^2) = 5*(1-0.64) = 5*0.36 = 1.8$ unique modes. If in fact, the two warnings were presented in different modes, the EM feature would be 1.8-2 = -0.2. Alternatively, if by chance the second warning was presented in the same mode as the first warning, then EM would be 1.8-1 = 0.8. Thus, negative values correspond to warnings occurring in different modes more often than expected, while positive corresponds to more consistency in modes across warnings than would be expected. This Expected Modes feature has the advantage of scaling appropriately based on the number of events that occurred (and are being considered).

Equation 2.1 Number of expected modes based on the number of warnings

$$EMR(\#W) = 5*(1- (0.8)^{\#W})$$

New Features Distribution

Figure 2.2 shows the feature distribution for the 2 new features using the Naïve group corresponding to a total of 150 outcomes for the 15 trials for each of the 10 subjects. Alternatively, Figure 2.3 shows the distribution of the second version of these features. As detailed in Table 2.3, the second version of these features captures consistency in modes over longer time periods. It shows how the second version of the initial features decreases considerably the number of 0 scores or provides more distinction to this score. The initial features showed more than half of the times a 0 score because more than half of the trials had a single warning.

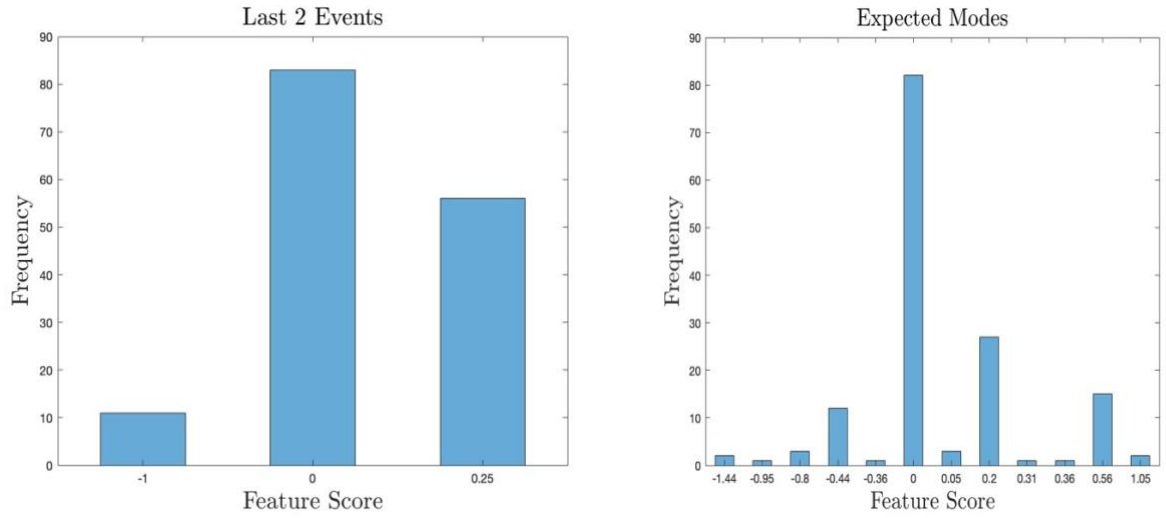

Figure 2.2 Histogram of scores for new initial features

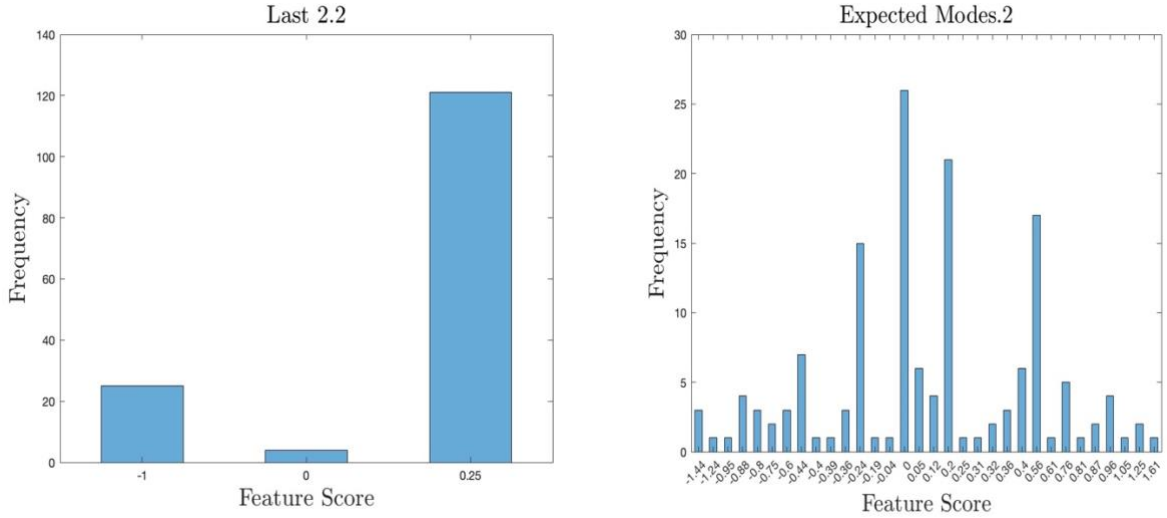

Figure 2.3 Histogram of scores for the second version of new features

Chapter 3: Results

New Features Performance

Figure 2.1 showed a decrease in performance on workload when developing Model 6 trained with Naïve data. However, this is the type of data that the intelligent adaptive autonomous system is expected to encounter. Therefore, we aim to develop a model that achieves a considerably lower mean absolute error when trained with Naïve data. Previous work focused on developing models using only non-adaptive data and did not incorporate any features that could capture the effects of user interaction with an adaptive autonomous system (Kintz et al., 2023; Rote et al., 2024). For this reason, we introduced new features, explained in the new features development section, designed to enhance predictive performance across all three cognitive states. While trust and workload performed better for model 6 with Naïve data compared to Static data, we wanted to explore whether these new features could yield even better performance. The new features were added one at a time to Model 6 and Figure 3.1 shows the mean absolute error across 100 Monte Carlo cross-validations. These models were trained using Naïve data with a split of 7 subjects for training and 3 for testing.

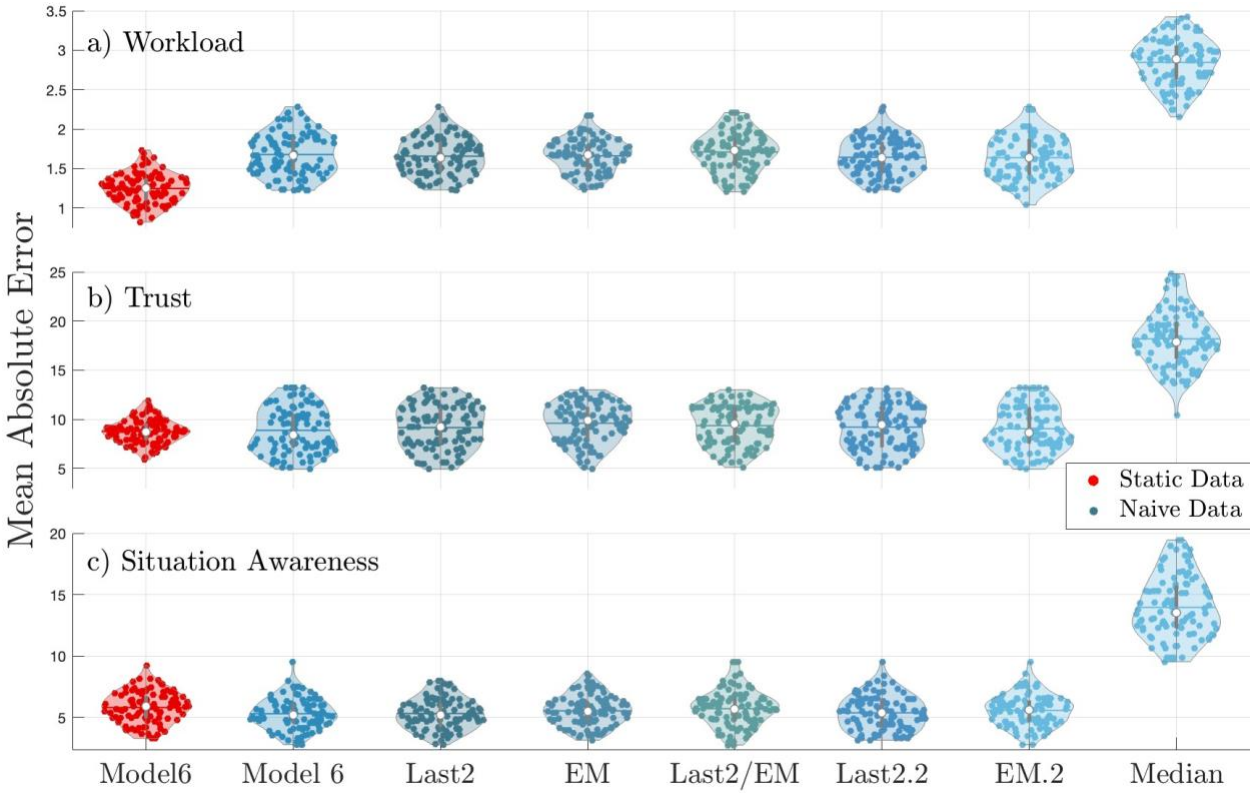

Figure 3.1 Performance comparison between Model 6 with Static and Naïve dataset vs Model 6 using Naïve data and allowing the model builder to select the new features defined in Table 2.3. For example, "Last 2", allows the model builder to select any of the features available to Model 6, but also includes the novel Last 2 feature to potentially be selected. The panels and formatting are the same as explained in Figure 2.1.

Overall, the addition of new features had a small effect on the model performance in comparison with the original Model 6 trained on Naïve data. For comparison the mean MAE for the models of workload with these features were Model 6: 1.68, Last2: 1.66, Last2.2: 1.64, EM: 1.66, EM.2: 1.64, Last2+EM: 1.71, all of which were much lower (i.e., better accuracy) than Median: 2.85 (Baseline), but somewhat worse than Model 6 trained with Static data: 1.25 (in red). Similarly, for comparison the mean MAE for the models of trust with these features were Model 6: 8.88, Last2: 9.16, Last2.2: 9.22, EM:9.61, EM.2: 9.05, Last2+EM: 9.37, Median: 18.19 (Baseline),

Model 6: 8.77 (trained with Static data). Similarly, for comparison the mean MAE for the models of situation awareness with these features were Model 6: 5.30, Last2: 5.32, Last2.2: 5.38, EM: 5.55, EM.2: 5.57 Last2+EM: 5.72, Median: 13.94 (Baseline), but slightly better than Model 6 trained with Static data: 5.81 (in red).

Figure 3.1 shows that all the new modes trained with the Naïve data and the new features achieve high performance (with a similar distribution) compared to the baseline of always predicting the median of the scale. However, for trust and situation awareness, the addition of the new features did not improve performance compared to Model 6 trained with Naïve data. For situation awareness, the lowest mean MAE corresponds to Model 6 trained with Naïve data, even surpassing Model 6 trained with Static data.

In the case of workload, the mean MAE increased considerably from Model 6 trained with Static data to Model 6 trained with Naïve data. Here, the models with Expected Modes.2 and Last2.2 achieved better performance, but the difference is so small (0.02) that it is likely not meaningful and not worth the addition of any new features. Implementing them would require more code to track the system mode across trials and additional real-time computation. Moreover, this small difference could potentially be reduced or disappear by performing more MCCVs. However, there was a big limitation in introducing the new features. They were only selected about 15% of the time, accounting for interactions with other features. This indicates there wasn't a considerable difference between Model 6 and the models incorporating

these new features, as shown by their very similar mean MAEs. This behavior was experienced across all cognitive states.

LASSO vs BIC vs AIC

Previous modeling methods for WTSA include exhaustive search followed by stepwise regression optimized by lowering BIC or alternatively using LASSO for feature selection. The exhaustive + stepwise approach has been consistently used in our laboratory due to its advantages, such as flexibility in adjusting the split between training and testing, compatibility with any data size, the ability to limit the number of predictors selected, generally much faster computational times with low or moderate number of potential features (as is the case here) and the option to force certain features to always be selected (Kintz, 2024). At the same time, relaxed LASSO has been favored for its ability to fit simultaneous models, along with reduced computation time and improved model quality as the number of predictors increases (Buchner et al., 2022; Buchner et al., 2025). LASSO also cannot be directly applied to predicting our Workload scores, because they are on an ordinal scale, and LASSO uses simple linear regression (rather than ordinal regression). Exhaustive + stepwise regression optimized by lowering AIC can be easily applied to the current code for exhaustive + stepwise regression optimized by lowering BIC, offering the possibility of creating models with additional features and interactions. Optimizing for AIC theoretically may improve predictive performance in terms of the accuracy of unseen observations, like we are striving for here. Thus, we aimed to compare two model-building methods previously applied to cognitive state estimation (BIC and LASSO)

to this novel method (AIC) which has not yet been employed for cognitive state estimation.

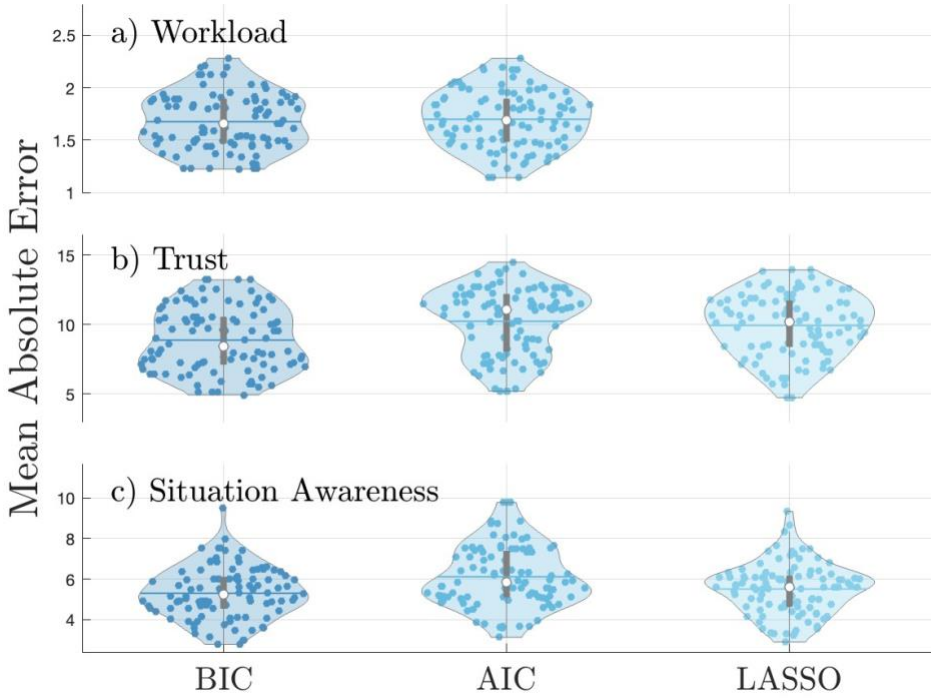

Figure 3.2 Performance comparison of models trained with Naïve data using the features allowed by Model 6 using different model-building methods. First, an exhaustive search followed by stepwise regression under Bayesian Information Criteria (BIC). Second, an exhaustive search followed by stepwise regression under Akaike Information Criteria (AIC). Third, using the Least Absolute Shrinkage and Selection Operator (LASSO). Workload has an ordinal scale that cannot be implemented under LASSO, which uses simple linear regression. See Figure 2.1 for the format.

Overall, all modeling methods performed similarly. For comparison the mean MAE for the models of workload were BIC: 1.68, and AIC: 1.70. It does not include a mean for LASSO because we treat workload as an ordinal scale that cannot be used in LASSO. Similarly, for comparison the mean MAE for the models of trust were BIC:

8.85, AIC: 10.24, and LASSO: 9.92. Similarly, for comparison the mean MAE for the models of situation awareness were BIC: 5.30, AIC: 6.11, and LASSO: 5.51.

The models trained with exhaustive search followed by stepwise regression under BIC demonstrated the best performance across all three cognitive states. They had the lowest mean MAE, the smallest range (difference between minimum and maximum MAE across 100 MCCVs), and considerably lower processing time compared to LASSO. While models with LASSO achieved similar performance with a good distribution, this method requires more computational power and time. This means that the BIC method can deliver similar or even better results compared to LASSO, but in less time, especially since the number of predictors for Model 6 is considerably lower than the number of observations.

Data Type and Size

We have three 3 different data cohorts of different sizes, Naïve (N=10), Static (N=14), and Both (Static + Naïve N=24). We wanted to explore the effect of the data size, which also investigated if it would be reasonable (and helpful) to merge the two datasets (Both = Static + Naive), despite the autonomous mode changing within trials vs. not potentially altering relationships between predictors and cognitive states.

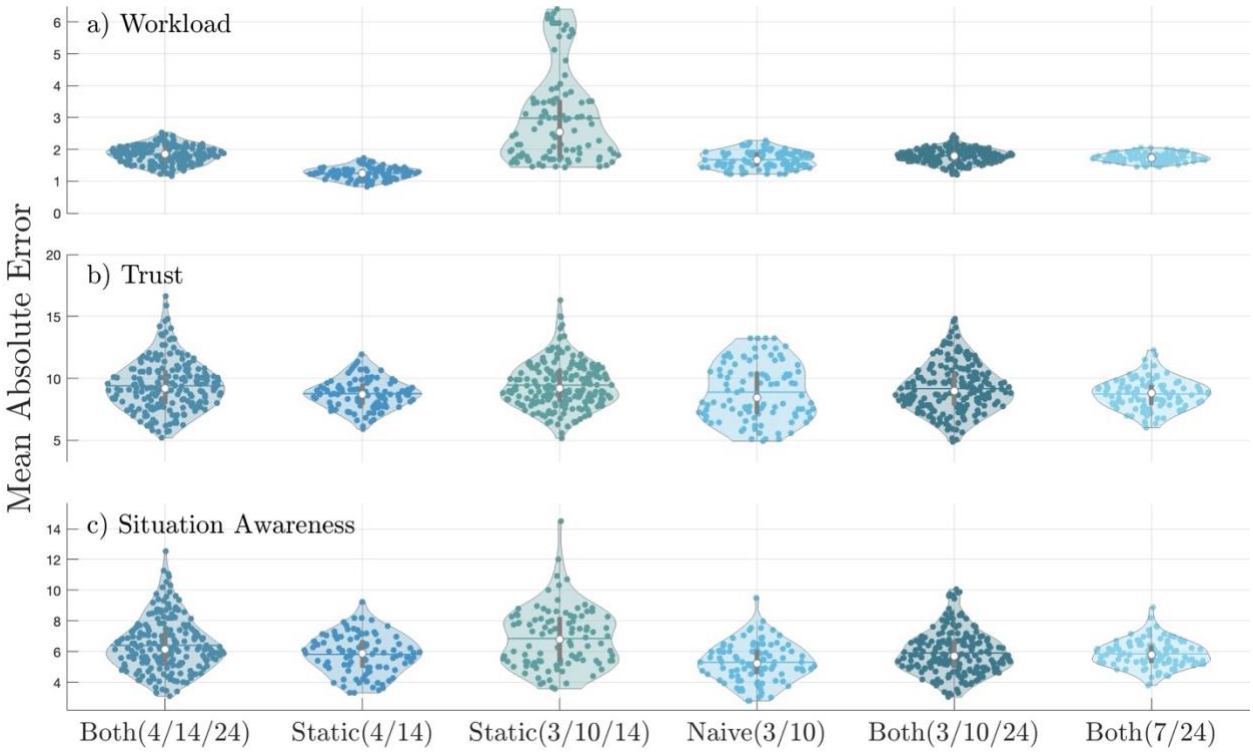

Figure 3.3 Performance comparison of Model 6 trained with different datasets (Static, Naïve, or Both (Static + Naive) and sizes. The naming convention on the x-axis corresponds to: (# of subjects for testing/data size feed to model / data size before downselection). If only 2 values are in parentheses, it means that the data size was not downselected. Starting on the left, the first 2 correspond to models trained with 14 subjects, trained with 10, and tested on 4. The first one corresponds to models trained using both data types but downselected to 14 before feeding the model to match the data size of the Static data. The third to fifth corresponds to a direct comparison between all data types with a training data size of 10 with a split of 7 training and 3 testing. This required downselecting Both from 24 to 10, and Static downselected from 14 to 10. The last one on the far right corresponds to a model trained with Both data types with a split of 17 subjects for training and 7 for testing (i.e., using all the available data, but roughly mimicking the train/test split proportions). See Figure 2.1 for the format.

The downselection of the data across all data types (Static, Naïve, and Both) led to a significant decline in model performance, accompanied by an increase in mean, range, and a shift in distribution toward the higher end. This was particularly

evident in MCCVs with very high MAE compared to MCCVs produced by the full dataset. The first comparison is between Both and Static with a data size of 14, with a split of training on 10 subjects and testing on 4. For workload, Both (4/14/24): 1.84, Static (4/14): 1.25. For trust, Both (4/14/24): 9.40, Static (4/14): 8.77. For situation awareness, Both (4/14/24): 6.40, Static (4/14): 5.81.

The second comparison is between Both, Static, and Naïve with a data size of 10, with a split of training on 7 subjects and testing on 3. For workload, Static (3/10/14): 2.98, Naïve (3/10): 1.68, Both (3/10/24): 1.79. For trust, Static (3/10/14): 9.45, Naïve (3/10): 8.88, Both (3/10/24): 9.18. For situation awareness, Static (3/10/14): 6.85, Naïve (3/10): 5.30, Both (3/10/24): 5.90. Lastly, the combination of Naïve and Static datasets was evaluated with a training split of 17 subjects and testing on 7. For workload (7/24): 1.74. For trust (7/24): 8.75. For situation awareness (7/24): 5.85.

For workload, the results mirrored previous findings, with the model trained on Static data achieving the highest performance. The second-best performance was observed in the model trained on Naïve data, followed by the model trained on Both without downselection, which exhibited a smaller range and a distribution closer to the mean. For trust, the model trained on Static data again demonstrated the best performance, followed by Both without downselection, and then the Naïve model (very small difference between Static and Both). In contrast, for situation awareness, the model trained on Naïve data performed best, followed by the Static model, and then Both without downselection. Although the optimal data type varied across the

three cognitive states, both without downselection consistently ranked among the top-performing models, exhibiting only slight differences in mean performance compared to the best model. It also had the lowest range across all datasets and cognitive states, maintained a distribution toward the mean, and included Naïve data that would be encountered when testing with previously unseen subjects.

Final Model

We found that the model with the lowest mean absolute error and the narrowest range between minimum and maximum values was the one trained using both Naïve and Static data (i.e., Both) with exhaustive search, where the feature selected was done by stepwise regression optimized by BIC, without including any additional features beyond those defined by Model 6. We note that some other models performed similarly well, but certainly combining both Naive and Static data to increase the size of the training (and test) datasets was beneficial. Otherwise, stepwise using BIC (e.g., rather than LASSO) allows for more rapid fitting and not including extra features intended to capture changes in mode from warning to warning simplifies the models and how they are built with negligible impact upon predictive accuracy. Table 3.1 presents all the features and interactions selected by the model for the three cognitive states, along with their corresponding coefficients, when trained using the full 24-subject dataset (i.e., no train/test split).

Table 3.1: WTSA model coefficients for the model with the best performance

Feature	Workload	Trust	Situation Awareness
Intercept	-	69.789	36.129
Level of Autonomy	0.127	0.112	-0.432
Level of Transparency	-0.047	-0.585	-0.621
Number of Warnings in a Trial	-0.703	-	-4.111
Number of Rejected Options	-	-5.651	-
Fraction of Time Spent in Secondary Task	2.325	-5.588	-15.906
Time Since Last Warning	0.018	-	-
Number of Tab Switches	0.134	0.342	0.519
Level of Transparency: Selected Accept and Reject in Same Trial	-1.249	-	-
Number of Warnings in a Trial: Fraction of Time Spent in Secondary Task	-	-	4.238

This model provides the best performance in predicting workload, trust, and situation awareness for unseen subjects. However, note that the coefficients on the Level of Autonomy and Level of Transparency were quite small. This suggests that an adaptive autonomous system that aims to alter an operator's cognitive states (e.g., decrease workload when it is too high) is unlikely to have much capability to modify the user's cognitive states.

Thus, as a final contribution, we performed an analysis where we aimed to predict each workload, trust, and situation awareness, but without any of the Simulation Events or Operator Actions features described in Table 2.2, but instead

only with the Autonomous System Modes. In the Static and Naive dataset, the Autonomous System Mode for each warning is determined by the experimenter and thus serves as the experimental independent variable (rather than an observational parameter that is either random or determined by the subject's decisions/actions). The purpose of this analysis was just to quantify the relative impact of a warning being presented in one autonomous system mode vs. another. Quantifying the magnitude of these effects is critical for future adaptive autonomous systems which aim to change modes in order to keep operator cognitive states at ideal levels. Specifically, if changing between the available autonomous system modes has negligible impact on operator cognitive states, the "lever" is too small for an adaptive autonomous system to be effective at manipulating those states and the adaptive system will be ineffective, even if it can non-disruptively estimate the operator's cognitive state very accurately (Kintz, 2024).

To better understand the impact of the autonomous system (i.e., its Level of Autonomy and Level of Transparency) on the operator's mental workload, we computed the ordinal regression model coefficients for low (-1) or high (1) levels of autonomy, relative to the baseline (0), and the corresponding values with low (0) or high (1) level of transparency. Note that these are on the ordinal regression scale and do not immediately map to modified Bedford workload scale values.

Table 3.2: Autonomous system effect on Workload

Level of Autonomy / Level of transparency	-1	0	1
0	-0.127	0	0.127
1	-0.174	-0.047	0.08

These results imply that the two aspects of the modes that an autonomous system could manipulate to guide the user toward an ideal cognitive state had a very slight impact on cognitive states. For example, the slightest difference required to shift between two levels of workload on the modified Bedford workload scale (e.g., from a 3 to a 4 on the ordinal scale), is 0.343, yet Table 3.2 shows that the maximum achievable change is only 0.301 (going from low level of autonomy but high transparency in the lower left at –0.174 to high autonomy and low transparency in the upper right at 0.127). This implies that changing the autonomous system mode is unlikely to be able to change the workload level by even one level in a systematic manner. Similarly, for trust and situation awareness, the different autonomous system modes were only able to accomplish a maximum linear change of -0.697 and -1.053 for the two scales with ranges of 72 and 60, respectively. The small coefficients directly associated with the autonomous system mode further suggest that the effect of the changing between different modes is minimal, indicating that the modes are highly similar in terms of systematic impact upon cognitive states. This implies that previously observed effects in Kintz 2024, were likely due to transitions between system modes rather than the individual impact of each mode on cognitive states.

Chapter 4: Conclusion

Summary

Aim 1: New features that consider the autonomous system's consistency within and across trials did not improve model predictive accuracy, given the variations between each MCCV train/test split. Although models with these features showed a slightly lower mean MAE, the added complexity of estimating these features in real time outweighed the performance gain. Moreover, the performance difference was so small that additional Monte Carlo cross-validations could cause this difference to be reduced.

Aim 2: Models trained with both Static and Naïve Data without downselection consistently ranked among the top-performing models across cognitive states, exhibiting only slight differences in mean performance compared to the model with the lowest mean. At the same time, they also demonstrated the smallest range between the minimum and maximum MAE across MCCVs. This finding suggests that including more data allows the model to have better performance when predicting unseen data (even if pooling data across Static and Naive datasets).

Aim 3: No considerable differences in predictive accuracy were observed when building models with exhaustive search followed by stepwise regression optimized by either BIC or AIC or when using LASSO. This suggests that while feature selection is necessary, the mechanism of performing feature selection may not be critical for enhancing model predictive accuracy. Since stepwise regression is less

computationally expensive than LASSO and optimizing for BIC is consistent with what has been used previously [24,43], this is a reasonable model-building approach to use going forward.

Limitations

The main limitation when developing new features to account for the autonomous system changing modes is that around 55% of trials include only one warning (and thus one autonomous system mode). Even when considering trials where the autonomous system's suggestion was wrong and generated an extra warning, or when the user rejected a correct suggestion (triggering a new warning after 20 seconds), these new features associated with different autonomous modes between warnings only carry value in fewer than half of the trials, making them difficult for the model builder to leverage effectively.

Another limitation is that LASSO cannot be used for the ordinal scale (Workload). Results showed that Workload is the cognitive state with the largest differences in predictive accuracy (MAE) between different types of models, yet we were unable to compare it when trained with a different model builder. Additionally, Model 6 substantially reduced the number of features used for model building, which limited the potential for interactions between the new features and other features that could capture the effect of the adaptive autonomous system.

Future Work

Results have shown that interacting with an adaptive autonomous system has an impact on cognitive states. Future work will explore developing new features to capture this impact like a feature that evaluates the consistency of the system over all the events that happened during the last 2 minutes. In prior data collection (Kintz et al., 2023), to ensure the autonomous system provided correct vs. incorrect suggestions at a given rate, all suggestions on a given trial were programmed to be either correct or incorrect. Going forward we will consider the effects of a system in which the correctness of the system varies from warning to warning instead of a fixed state throughout the trial. How are cognitive states affected when the system has an opportunity for redemption? Will the user perceive that the system is learning from its mistakes, and how will cognitive states change if the system repeats the same error?

At the same time, the results demonstrated that Model 6 optimized by BIC and trained with both Static and Naïve datasets have a fairly accurate performance when predicting unseen data. However, the coefficients for the autonomous system mode are not large enough for the system to effectively guide the user toward the ideal range of cognitive states. Future work will explore developing new modes that are more distinct from one another and focused on having a more substantial impact on a single cognitive state at a time. These modes, which we have begun to implement, are described below.

A mode focused on decreasing workload can be split into two stages of communication. First, an initial vague warning informs the user of the anomaly without offering additional details about what actions could be taken, what might be the cause of the anomaly, or the effect of taking potential actions. This warning includes a 10-second timer to prompt a decision: accept, reject, or wait for further information. It encourages or enables the user to make a quick choice rather than requiring a detailed investigation across all systems. If the timer expires, a second warning appears after 5 seconds, providing more details about the problem, a possible solution, and a brief explanation of why the system recommends taking a particular action. Since this mode aims to reduce workload, it includes a 20-second timer; if the user doesn't accept or reject when the timer ends, the system will proceed as if the user decided to accept. This approach is expected to considerably reduce workload, especially if the user responds to the first message, although it may lead to a considerable decrease in situation awareness if the initial message is accepted or rejected as the user is not able to have any information about the anomaly. Trust in this mode may be expected to be dynamic, largely depending on whether the system's action successfully resolves the problem (Manzey et al., 2012; Yang et al., 2021).

A mode focused on increasing workload would require the user to identify both the problem and the solution. It begins with an initial warning message stating that an error was detected, and that corrective action is required. The system then guides the user through a step-by-step process, with feedback at each stage for the system to provide the next step. Once the anomaly is detected, the system provides three

possible solutions along with a reject option. In most cases, we noticed that users spend very little time in the decision-making process and tend to accept the recommendation even when the system may be wrong. This mode is expected to increase both workload and situation awareness, as the user must review all metrics and devices, take actions, report to the system, and carefully consider which action to take (Endsley & Kaber, 1999). It is expected that this mode decreases trust in most cases due to the lack of support and guidance during decision-making.

A mode focused on situation awareness should help the user understand why a warning appears, why they should (or should not) accept the suggested action and provide visual support of what the action will do. In this approach, the system provides a detailed message that includes the tab where the anomaly was detected, the current level, and the nominal ranges, followed by an explanation for accepting the action. Simultaneously, the system offers visual support by highlighting the section of the table corresponding to the value bounds and indicating which devices will be activated if the action is taken. This mode is expected to considerably increase situation awareness and enhance the user's understanding of the effects of the actions (Adams et al., 1994). It could increase workload if the error involves multiple steps but is anticipated to have a moderate overall effect. Trust is expected to increase substantially due to the detailed explanation, visual support, and clear connection between the warning and the solution, which helps the user feel supported even if the system is occasionally wrong.

Lastly, a mode focused on increasing trust involves a "what if?" approach. The system presents a warning, the corrective action, and two exploratory options: “What if accept?” and “What if reject?” It then simulates the potential outcomes and reports the effects of accepting or rejecting. The user is limited to these exploratory options, able to select them only once, without having the option to accept or reject the solution. After the simulation, only the choices of accept or reject remain available. In this mode, the user can clearly see the consequences of the action, even when the system is wrong, which could tend all cognitive states toward the nominal range. Situation awareness would slightly increase, as the user will know exactly what would happen and how it would affect the system. Trust would increase because even when the system is wrong, the simulation always shows the correct outcome. Workload would experience a slight increase since the user is presented with two warnings, two decisions, and new information.

As the next steps, beyond the scope of this thesis, we plan to evaluate operator cognitive states using each of these new autonomous system modes, in a human subject experiment where the mode changes from warning to warning. This will inform future models that predict operator workload, trust, and situation awareness. Model building will be informed by the results of this thesis (i.e., using larger subject pools, considering but likely not requiring features to capture the change in autonomous system mode, and employing stepwise optimized for BIC, which performed well in our analysis. Informed by these non-disruptive models for estimating operator cognitive state, a future adaptive autonomous system

implementation can be developed to maintain cognitive states at ideal values and be evaluated in a human subject experiment.

Bibliography

Anderson, A. P., Clark, T. K., & Kong, Z. (2020). Adaptive autonomy for future spacecraft habitats. In ICSR 2020: Human-Robot Interaction for Space Robotics Workshop, Golden, CO, USA.

Adams M. J., Tenney Y. J., Pew R. W. (1994). Situation awareness and the cognitive management of complex systems. *Human Factors*, 37, 85–104.

Bechtold B. (2016). Violin plots for. Matlab. Github Project. https://github.com/bastibe/Violinplot-Matlab, https://doi.org/10.5281/zenodo.4559847

Borghetti, B. J., Giametta, J. J., & Rusnock, C. F. (2017). Assessing continuous operator workload with a hybrid scaffolded neuro ergonomic modeling approach. Human Factors, 59(1), 134–146. https://doi.org/10.1177/0018720816672308

Buchner S. L. (2022). *Multimodal feature selection to unobtrusively model trust, workload, and situation awareness*. [Master's thesis]. University of Colorado at Boulder.

Buchner, S. L, Kintz, J., Banerjee, N., Zhang, J., Clark, T.K., and Anderson A.P. (2025). "Assessing Physiological Signal Utility and Sensor Burden in Estimating Trust, Situation Awareness, and Mental Workload" Journal of Cognitive Engineering and Decision Making, doi: 10.1177/15553434241310084.

Byeon, S., Yuh, M., Choi, J., Jain, N., & Hwang, I. (2025). Workload classification for function allocations in human–autonomy teaming using noninvasive measurements. In AIAA SCITECH 2025 Forum. https://doi.org/10.2514/6.2025-2253

Calhoun, G., Ruff, H., Frost, E., Bowman, S., Bartik, J., & Behymer, K. (2021). Performance-based adaptive automation: Number of task types and response time measures triggering level of automation changes. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 65(1), 37–41. https://doi.org/10.1177/1071181321651099

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From "automation" to "autonomy": The importance of trust repair in human-machine interaction. Ergonomics, 61(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

Diamond, M., Leon, G. R., & de León, P. (2025). Mars mission communication delays and impacts on mission controller performance, workload, and stress. Aerospace Medicine and Human Performance, 96(1), 67–70. https://doi.org/10.3357/AMHP.6550.2025

Dzindolet, M., Pierce, L., Beck, H., & Dawe, L. (2002). The perceived utility of human and automated aids in a visual detection task. Human Factors, 44, 79–94. https://doi.org/10.1518/0018720024494856

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In Proceedings of the Human Factors Society Annual Meeting, 32(2), 97–101.

Endsley M. R., Kaber D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42(6), 462–492.

Richardson E. E., Buchner S. L., Kintz J. R., Clark T. K., 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* (pp. 2438-2440). https://dl.acm.org/doi/10.5555/3635637.3663186

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. Human Factors, 54(6), 1008–1024. https://doi.org/10.1177/0018720812443983

Frank, J. D., McGuire, K., Moses, H. R., & Stephenson, J. (2016). Developing decision aids to enable human spaceflight autonomy. AI Magazine, 37(4). https://doi.org/10.1609/aimag.v37i4.2683

Frasheri, M., Cürüklü, B., Esktröm, M., & Papadopoulos, A. V. (2018). Adaptive autonomy in a search and rescue scenario. In 2018 IEEE 12th International Conference on Self-Adaptive and Self-Organizing Systems (SASO) (pp. 150–155). IEEE. https://doi.org/10.1109/SASO.2018.00026

Guo, Y., & Yang, X. J. (2021). Modeling and predicting trust dynamics in human-robot teaming: A Bayesian inference approach. International Journal of Social Robotics, 13(8), 1899–1909. https://doi.org/10.1007/s12369-020-00703-3

Hancock, P. A., & Matthews, G. (2018). Workload and performance: Associations, insensitivities, and dissociations. Human Factors. https://doi.org/10.1177/0018720818809590

Hancock, P. A., Jagacinski, R. J., Parasuraman, R., Wickens, C. D., Wilson, G. F., & Kaber, D. B. (2013). Human-automation interaction research: Past, present, and future. Ergonomics in Design, 21(2), 9–14. https://doi.org/10.1177/1064804613477099

Harrivel, A. R., Stephens, C. L., Milletich, R. J., Heinich, C. M., Last, M. C., Napoli, N. J., Abraham, N. A., Prinzel, L. J., Motter, M. A., & Pope, A. T. (2017). *Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing* [Conference paper]. *AIAA SciTech Forum*, Grapevine, TX. https://doi.org/10.2514/6.2017-1135

Heard, J., & Adams, J. A. (2019). Multi-dimensional human workload assessment for supervisory human-machine teams. Journal of Cognitive Engineering and Decision Making, 13(3), 146–170. https://doi.org/10.1177/1555343419847906

Heard, J., Fortune, J., & Adams, J. A. (2020). SAHRTA: A supervisory-based adaptive human-robot teaming architecture. In Proceedings of the 2020 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), 1–8. https://doi.org/10.1109/CogSIMA49017.2020.9215996

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics, 4(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kintz, J. R. (2024). *An Adaptive Autonomous System Informed by Unobtrusive Measures of Trust, Mental Workload, and Situation Awareness for Deep Space Exploration* (Doctoral dissertation). University of Colorado Boulder, Boulder, CO, USA.

Kintz, J. R., Buchner, S. L., Anderson, A. P., & Clark, T. K. (2023). Predicting operator cognitive states for supervisory human–autonomy teaming. In 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 5146–5153. https://doi.org/10.1109/SMC53992.2023.10394254

Kintz, J. R., Shen, Y.-Y., Buchner, S. L., Anderson, A. P., & Clark, T. K. (2023b). A simulated air revitalization task to investigate remote operator human–autonomy

teaming with communication latency. In 52nd International Conference on Environmental Systems, 2023.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Luo, R., Weng, Y., Wang, Y., Jayakumar, P., Brudnak, M. J., Paul, V., Desaraju, V. R., Stein, J. L., Ersal, T., & Yang, X. J. (2021). *A workload adaptive haptic shared control scheme for semi-autonomous driving*. *Accident Analysis & Prevention, 152*, 105968. https://doi.org/10.1016/j.aap.2021.105968

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. Journal of Cognitive Engineering and Decision Making, 6(1), 57-87. https://doi.org/10.1177/1555343411433844

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs: Journal of Cognitive Engineering and Decision Making. https://doi.org/10.1518/155534308X284417

Parisi, M., Panontin, T., Wu, S.-C., McTigue, K., & Vera, A. (2023). Effects of communication delay on human spaceflight missions. In 14th International Conference on Applied Human Factors and Ergonomics (AHFE), San Francisco, CA, USA, July 20–24.

Pereira, E., Sigcha, L., Silva, E., Sampaio, A., Costa, N., & Costa, N. (2025). Capturing mental workload through physiological sensors in human-robot collaboration: A systematic literature review. Applied Sciences, 15(6), 3317. https://doi.org/10.3390/app15063317

Roscoe A. H., Ellis G. A. (1990). A subjective rating scale for assessing pilot workload in flight: A decade of practical use. Royal Aerospace Establishment Farnborough (United Kingdom). https://apps.dtic.mil/sti/citations/ADA227864

Rote, N. C., Kintz, J. R., Richardson, E. E., Hayman, A. P., & Clark, T. K. (2024). Improving predictions of cognitive states for an adaptive autonomous system. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 68(1), 1585–1590. https://doi.org/10.1177/10711813241277522

Rollock, A. E., & Klaus, D. M. (2022). Defining and characterizing self-awareness and self-sufficiency for deep space habitats. Acta Astronautica, 198, 366–375. https://doi.org/10.1016/j.actaastro.2022.06.002

Schwarz, J., & Fuchs, S. (2018). Validating a "Real-Time Assessment of Multidimensional User State" (RASMUS) for Adaptive Human-Computer Interaction. 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 704–709. https://doi.org/10.1109/SMC.2018.00128

Selcon, S. J., & Taylor, R. (1990). Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design. In Situational Awareness in Aerospace Operations (AGARD-CP-478), 478, 1–8.

Stanton, N. A., Chambers, P. R. G., & Piggott, J. (2001). Situational awareness and safety. Safety Science, 39(3), 189–204. https://doi.org/10.1016/S0925-7535(01)00010-8

Wickens C. D., Dixon S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Science, 8(3), 201–212. https://doi.org/10.1080/14639220500370105

Wischnewski, M., Krämer, N., & Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Article No. 755, pp. 1–16). ACM. https://doi.org/10.1145/3544548.3581197

Yang, X. J., Schemanske, C., & Searle, C. (2021). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. Human Factors. https://doi.org/10.1177/00187208211034716

Appendix A: New Features Model

We develop a new model for the three cognitive states using all the available Naïve data. It means the model trained with 10 participants. Also, the model building had the option to select all the features from Model 6 and the four new features we developed.

Table 3.1: WTSA model coefficients for the model trained with Naive data and all the new features available

Feature	Workload	Trust	Situation Awareness
Intercept	-	67.687	34.044
Level of Autonomy	0.013	-1.717	-0.032
Level of Transparency	0.118	-2.835	-2.073
Number of Rejected Options	-	-5.088	-
Level of Autonomy: Level of Transparency	-	7.7193	-
Number of Warnings in a Trial	-	-	-1.772
Time Since Last Warning	0.027	-	-
Fraction of Time Spent in Secondary Task: Number of Tab Switches	0.217	-	-
Number of Warnings in a Trial: Number of Tab Switches	-0.044	-	-
Level of Autonomy: Switched Tabs After Warning and Before Confirm Action	1.417	-	-
Time Since Last Checked a Tab: Expected Modes	-0.091	-	-
Selected Accept and Reject in Same Trial: Last 2	2.513	-	-

Level of Autonomy: Selected Accept and Reject in Same Trial	-1.417	-	-

Table 3.1 shows the coefficient, and the features selected when training a model with the Naïve data and allowing for all the new features to be selected as well as its interactions. The results are very similar compared with the best model for trust and situation awareness. This time fewer features got selected, the new features never got selected, and there is not an increase in the coefficients on the level of autonomy or transparency. The coefficient for the Level of autonomy for trust became negative in this model. Suggesting that in this model a higher system autonomy represents lower trust, but the effect is very small due to its low coefficient. In the case of workload, the initial version of the new features got selected as interaction with another feature. In this case, Level of Transparency had a higher coefficient with the sign flipped in the sign now making a direct relationship between the system providing explanation and workload. The coefficient for Level of Autonomy had a considerable decrease, making its impact almost negligible. Overall, these results confirm previous behaviors seen when training models with the new features. The new features were only selected between 10 and 20 percent across 100 MCCVs, suggesting that they don't have a meaningful impact on the cognitive states to be included in the model. As shown by the results in Table 3.1, the new features didn't get selected as a feature by itself and were only selected as interactions with other features. But when comparing these interactions, we can see that it got selected with

selected accept and reject in the same trial which only had a meaningful value for about 10 percent of all the trials. The second interaction was with time since last checked a tab and this feature always had a very small value and the coefficient for this interaction is so small that its effect is reduced in most of the cases. Overall, the new features seem to not have a considerable impact on the cognitive states.