Dynamic Estimation of Mental Workload and Operator Accuracy for Time-Constrained Binary Classification Tasks

Raihan Seraj, Aditya Mahajan and Jerome Le Ny

***Abstract*—Human cognitive states, such as mental workload, play a pivotal role in decision making processes within human automation teams. Although subjective measures of mental workload can be obtained using standard questionnaires like the NASA-TLX, their administration is often impractical as it interferes with the primary tasks of the human operator. Therefore, it is of interest to estimate these subjective measures from less intrusive observations. Evidence suggests that mental workload is a dynamic process so incorporating historical measurements could reduce its estimation error. Additionally, the estimation of operator performance in human automation teams is essential in optimizing task effectiveness and facilitating efficient resource allocation. In this work, we consider a scenario where a human and an automation solve binary classification tasks under time constraints. We present and compare different dynamic schemes to estimate the operator's performance, i.e., classification accuracy, and her subjective ratings on subscales of the NASA-TLX questionnaire, which measure mental workload across multiple dimensions. These schemes differ in the information available for estimation. We test these schemes on data collected from a scenario where a human and an automation perform a series of classification tasks for simulated mobile objects. Our analysis of the interaction data and the estimation schemes indicates that employing dynamic estimation for certain NASA-TLX subscale ratings leads to decreased estimation errors.**

***Index Terms*—Human Automation Interaction, Mental Workload, Human Performance.**

I. Introduction

AUTONOMOUS systems are becoming pervasive and impact applications in manufacturing, autonomous driving systems, disaster management, or healthcare. A crucial element of these systems is the capacity of the human operators to make sound decisions, which is inherently influenced by cognitive factors such as mental workload (MW), which cannot be directly observed in real-time but impacts task efficiency, system effectiveness, and overall operational outcomes. Maintaining an appropriate level of MW is critical, as insufficient MW can lead to disengagement and excessive MW leads to heightened fatigue, both of which can deteriorate performance in human-automation teams [1], [2]. Therefore, assessing MW is essential for designing appropriate adaptive automation interfaces [3], [4].

A frequent means of measuring MW is by collecting subjective responses to standard questionnaires [5] such as the NASA-TLX (Task Load Index) [6], which evaluates workload across multiple subscales, encompassing mental demand, physical demand, temporal demand, performance, effort and frustration. Although subjective ratings have a high face validity [5], collecting them is intrusive [7] and forces the operators to disengage from their primary tasks. An alternative approach to assess MW is to use *passive measurements*, combined with certain *task characteristics* (e.g., task complexity, number of subtasks, etc.). Passive measurements include i) physiological measurements such as heart rate, galvanic skin response, etc. [8]; and ii) operator performance metrics (OPMs) in primary or secondary tasks, such as accuracy or error rates, response times, etc. [9], [10].

MW models based on physiological measurements such as electrocardiograms, event related potentials, electroencephalograms and galvanic skin responses are studied in [11]–[13], [13]–[15]. See [16] for a survey. MW models based on OPMs in primary or secondary tasks include [17]–[21] and use tools such as linear regression [17], petri nets [18], or hidden Markov models [19]–[21]. Physiological measurements often require costly sensors and controlled conditions, which restricts their use [22]. Therefore, we use only OPMs as passive measurements to assess mental workload.

For many tasks, we can identify certain primary OPMs of importance (e.g., operator accuracy or error rates in a classification task), which could also be estimated using secondary OPMs (operator response times, efficiency of the operator's interactions with the interface, etc.), task characteristics (complexity, task load or intensity, etc.), and possibly past values of primary and secondary OPMs. Indeed, there may be situations where direct measurements of primary OPMs are unavailable, e.g., for classification tasks where the ground truth is not known. In such cases, identifying reliable indirect indicators of performance becomes particularly important.

Given the intuitive correlation between MW and operator performance [23], there is also an interest in leveraging subjective MW measurements to estimate primary OPMs and predict future changes in performance. Linear regression models are developed in [24] to establish quantitative rela-

R. Seraj and A. Mahajan are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0E9, Canada. Emails: raihan.seraj@mail.mcgill,ca, aditya.mahajan@mcgill.ca.

J. Le Ny is with the Department of Electrical Enginieering, Polytechnique Montreal, Montreal, QC H3T 1J4, Canada. Email: jerome.le-ny@polymtl.ca

This research was supported by the Innovation for Defence Excellence and Security (IDEaS) Program of the Canadian Department of National Defence through grant CFPMN2-037 and Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2021-03511 and RGPIN-2018-5287. The experiment protocol with human subjects in this study has been reviewed by McGill Research Ethics Board (REB), approval number: 22-02-017. The authors thank Rexys, Inc for helping in the design and development of the SCSS simulator that was used for the online experiments.

tionships between a primary OPM (flight technical errors) and NASA-TLX ratings of the performance, physical demand and temporal demand subscales. However, because these ratings were collected after each scenario where performance was measured, this study is not targeted at real-time performance estimation. In [25], a long short-term memory (LSTM) neural network [26] processing the outputs of a multi-dimensional workload assessment algorithm is used to predict operator performance up to 5 minutes in the future on an aircraft supervision task.

It is well established in the literature that cognitive states such as MW are dynamic variables with temporal correlations [21], [27]. The same holds true for some OPMs [28]. So, incorporating historical measurements of MW, primary and secondary OPMs, and task characteristics in the estimation models may help reduce estimation error. This dynamic aspect of model estimation has not received sufficient attention in the literature. A few studies use HMMs to estimate cognitive models [19]–[21], but they do not compare with other dynamic estimation techniques. Some mathematical models accounting for the temporal correlation of MW utilize dynamic queue formulations [29], [30], but these models are theoretical and have not been verified experimentally. A related concern is that collecting subjective ratings to questionnaires is an intrusive process, so that it can only be performed infrequently, which can result in increased estimation errors. Hence, the trade-off between the frequency at which subjective MW measurements are collected and estimation quality needs to be investigated.

II. Research Questions and Contributions

The previous discussion motivates the following research questions:

- Q1 Does including the history of passive measurements and task characteristics help in reducing estimation error of MW (question Q1a) and primary OPM (question Q1b)?
- Q2 Does adding past subjective MW measurements to the history of passive measurements and task characteristics help in reducing estimation error of MW (question Q2a) and primary OPM (question Q2b)?
- Q3 How does the estimation error of MW (question Q3a) or primary OPM (question Q3b) vary with the frequency with which the operators are asked to provide subjective MW measurements?
- Q4 How does the estimation error of primary OPM vary when past primary OPM measurements are not available?

Given the practical importance of the NASA-TLX subscale ratings in assessing MW [24], [31], in this paper we equate estimating MW with estimating these subjective ratings, as in [24] for instance. The original NASA-TLX paper recommends combining all the reported scores into a single score. Later papers argue that such a combined score is not very meaningful [32]. So we separately estimate each subscale ratings and consider them as different facets of MW.

To address the research questions above, we consider repeated binary classification tasks where human operators interact with a simulated environment closely inspired by the Simulated Combat Control System (SCCS) microworld of [33], which has been used in previous research to explore situation awareness and task performance within dynamic decision-making settings relevant to command and control operations. We design an experiment where the operators observe the parameters of different aircraft and classify them as hostile or non-hostile by following a pre-specified decision tree. They then answer the standard NASA-TLX [6] questionnaire to provide subjective measurements of different factors associated with their MW. For our experimental setup we consider the following metrics: i) operator classification accuracy (as primary OPM), ii) operator average classification time (as secondary OPM), iii) number of tasks given to the operator and the round in which the tasks are performed (as task characteristics). Details of the experimental setup and the metrics are presented in Section III. Based on the interaction data obtained from human participants, we consider the following estimation schemes to answer our research questions (see also Table II for more details):

- S1 Estimate MW (scheme S1a) or primary OPM (operator classification accuracy, scheme S1b) based on the last available values of primary and secondary OPM measurements and task characteristics.
- S2 Estimate MW (scheme S2a) or primary OPM (scheme S2b) based on the history of primary and secondary OPM measurements and task characteristics.
- S3 Estimate MW (scheme S3a) or primary OPM (scheme S3b) based on the history of primary and secondary OPM measurements and task characteristics and the last available NASA-TLX subscale ratings.
- S4 Estimate MW (scheme S4a) and primary OPM (scheme S4b) based on the history of primary and secondary OPM measurements and task characteristics and the history of NASA-TLX subscale ratings.
- S5 Estimate MW (scheme S5a) and primary OPM (scheme S5b) based on the history of primary and secondary OPM measurements and task characteristics and a *periodic subset* of the past NASA-TLX subscale ratings.

We also consider variations of S1b-S5b, denoted S1b'–S5b', where past measurements of primary OPM are not available to estimate current primary OPM. Since we are not aware of any existing study comparing the performance of different estimators for predicting MW and OPM, we have opted to evaluate multiple state-of-the-art estimators, and, for each estimator, select the one with the best performance. This is helpful to avoid the answers to our research questions to depend on the performance of a specific estimator. We consider and present in Section IV the following estimators for the different schemes: i) random forest regressor ii) support vector regressor iii) XGBoost regressor iv) recurrent neural network with long short term memory (RNN-LSTM) v) Kalman filtering on transformed data vi) non linear filtering with input-output hidden Markov model (IO-HMM) vii) autoregressive integrated moving average with exogenous input (ARIMAX) on transformed data. We compare the root mean square error (RMSE) (referred to as estimation error) across the different schemes to answer the research questions Q1–Q4 described above. Our findings are presented in Section V.

The main contributions of our work are as follows:

- We present different dynamic estimators and illustrate their utility in estimating operator MW (as measured by different NASA-TLX subscale ratings), and estimating primary OPM (as measured by classification accuracy).
- We analyze and statistically verify (based on the collected data and best performing estimator) whether including the history of primary and secondary OPM measurements and task characteristics along with the history of NASA-TLX subscale ratings reduces the estimation error for the estimation of MW and primary OPM.
- We show how the frequency at which we can obtain NASA-TLX subscale ratings affects estimation errors of MW and primary OPM, and outline strategies for adjusting estimator inputs to address missing values.

III. Experimental Setup and Data Collection

A. Description of the simulator

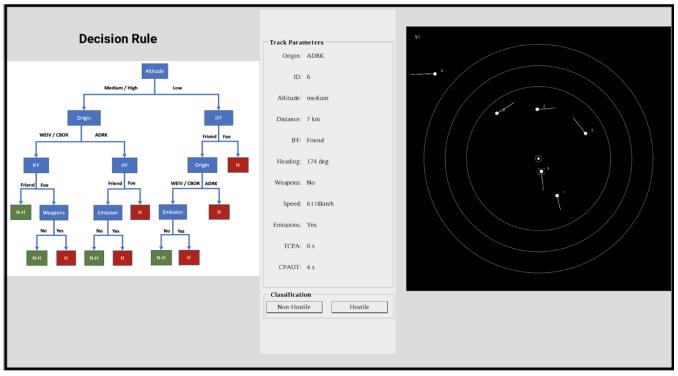

Fig. 1. The interface of SCCS microworld.

A between-subject experimental study was performed where participants interacted with a simulator closely inspired by the Simulated Combat Control System (SCCS) microworld described in [33]. This simulation environment represents a simplified naval anti-air warfare scenario. Participants play the role of an operator and monitor parameters of different aircraft appearing on a radar screen to classify them as hostile or non-hostile, by following a provided decision tree.

The simulator interface, shown in Figure 1, consists of three panels. The right panel shows a mock-up of a radar screen with aircraft represented as white dots and their direction of motion shown as a thin white line. The operator can select an aircraft by clicking on the white dot. Each aircraft has multiple parameters (origin, altitude, weapons, emissions, etc.) as shown in the middle panel. The bottom of the middle panel has two buttons to classify the selected aircraft as 'Hostile' or 'Non-Hostile'. When an aircraft is classified as 'Hostile', the color of its dot changes to red, when it is classified as 'Non-Hostile' the color of its dot changes to green. It is possible to re-select a previously classified aircraft and change its classification. The left panel shows a decision tree that the operator is asked to follow to classify the aircraft as 'Hostile' or 'Non-Hostile'. The top left corner of the radar screen shows a timer displaying the time (in seconds) remaining in the current round. Some additional details of the simulator, including the decision tree and aircraft parameters are shown in the supplementary material.

B. Description of the experiment

To minimize the learning effects [34], the participants started with 3 practice rounds of 120 seconds each to become familiar with the interface. The main experiment then consisted of 25 rounds of 120 seconds each. The participants had an option to take a break after every 5 rounds. To understand the impact of taskload on MW, the participants were shown in each round a subset of 30 aircraft, according to the following schedule:[1] a low tasload of 14 aircraft per round was shown in rounds $1-5$, $11-13$, and $18-20$; a medium taskload of 18 aircraft per round was shown in rounds $6-10$ and $23-25$; and a high taskload of 22 aircraft per round were shown in rounds $14-17$ and $21-22$. The remaining aircraft were not shown on the screen and automatically classified by the automation.

The participants were asked to classify each aircraft with the provided decision tree by starting from the top node and traversing the tree using the values of the parameters of each aircraft, until they reach a leaf node with the decision "Hostile" or "Non-Hostile". Participants were allowed to update their classification decisions multiple times and only their last classification decision was considered. At the end of each round, aircraft that were shown to the participants but left unclassified were randomly classified by the system with an accuracy of 50%. The aircraft not shown on the screen were classified by the automation using a different decision tree with fewer parameters.

At the end of each round, the classification accuracy of the participants and the automation were displayed on the screen. Then, the participants were asked to provide their subjective ratings on a scale of 1 to 7 for different subscales of the NASA-TLX questionnaire (screenshots showing how the classification information and questionnaire were displayed are provided in the supplementary material). Ratings for the physical demand subscale of the standard NASA-TLX questionnaire were not requested, as it is not relevant for the task.

C. Participants

A total of 26 participants (12 males, 14 females) with age ranging from 21 to 33 (mean 24.54, standard deviation 3.148) participated in this study. The data from two participants whose subjective ratings remained constant throughout the experiment was discarded. The participants were McGill University students recruited using online advertising about the research study and were compensated CAD 25 for participating. The experiment was conducted online and the participants provided their consent before starting. The study was approved by McGill University's Research Ethics Board (REB).

[1]There was no task randomization, i.e., each participant saw the same aircraft in the same order across the rounds.

TABLE I
LOGGED DATA FOR THE OPERATOR.

Logged data	Symbol	Explanation
Taskload	N_k	$\lvert\mathcal{C}_k\rvert$
Average classification time	τ_k	$\frac{1}{N_k}\sum_{i\in\mathcal{C}_k}\sum_{j\in\mathcal{I}_k} t^{i,j}_{\text{interval}}$
Operator classification accuracy	ACC_k	$\frac{\sum_{i\in\mathcal{C}_k}\mathbb{1}_{(c^i_k=g^i_k)}}{N_k}\times 100$
Operator ratings for each subscale	RES_k	Operator ratings for each NASA-TLX subscale, on a scale of 1 to 7

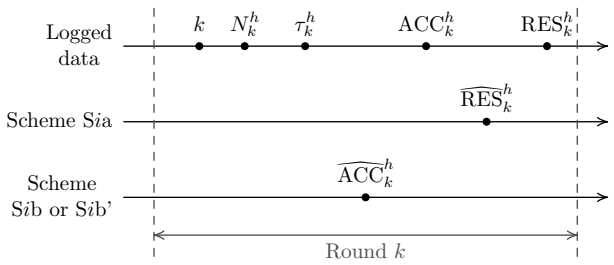

Fig. 2. Timing diagram showing when different variables are logged and when different estimates are generated.

D. Logged data

The simulator logs quantitative performance metrics. The details are shown in Table I. For each participant, $\mathcal{C}_k$ denotes the set of aircraft assigned to the operator at round $k \in \{1,\dots,25\}$. Its cardinality $|\mathcal{C}_k| = N_k$ represents the operator's taskload at round k. For each aircraft $i \in \mathcal{C}_k$, $\mathcal{I}^i_k$ denotes the number of times aircraft i was classified. For each $j \in \mathcal{I}^i_k$, $t^{i,j}_{\text{interval}} := t^{i,j}_{\text{classified}} - t^{i,j}_{\text{clicked}}$ denotes the time between when the i-th aircraft was clicked for the j-th time and when it was classified for the j-th time. Instances where an aircraft was clicked but not classified are not counted. For $i \in \mathcal{C}_k$, $g^i_k \in \{\text{hostile}, \text{non-hostile}\}$ denotes the ground truth for aircraft i and $c^i_k \in \{\text{hostile}, \text{non-hostile}\}$ denotes the final classification decision for aircraft i. Recall that aircraft not classified by the operator at the end of a round were randomly classified. Finally, we use the notation RES_k to refer to any of the 5 NASA-TLX subscale ratings collected from the operator at each round k (without differentiating in the notation between subscales, to streamline the presentation). Later, we also use the notation $\text{RES}^{\text{all}}_k$ to refer to the vector containing all 5 ratings collected at period k.

IV. METHODOLOGY

As discussed in Section II, we compare five classes of estimation schemes S1–S5 to answer questions Q1–Q3. The estimation schemes S1–S4 can be classified into two groups:

- Scheme S1 estimates MW or operator classification accuracy based only on the current observations. This estimation problem can be formulated as a supervised regression problem. Therefore, we use popular regression models, in particular random forest, XGBoost, and support vector regressors as estimators.
- Schemes S2, S3 and S4 estimate operator MW or operator classification accuracy based on the history of observations. For this, we use two types of dynamic models:
  (i) input-output models, where we use both regression models, with a finite window of past observations serving as features, and autoregressive integrated moving average with exogenous input (ARIMAX) time-series models;
  (ii) state-space models, where we use RNN-LSTM models, input-output hidden Markov models (IO-HMMs), and Kalman filtering.

Scheme S5 is similar to scheme S4 but with missing measurements. Some of the recursive filtering algorithms such as Kalman filtering and IO-HMMs are explicitly able to handle missing observations. For others, including regression models, RNN-LSTM, and ARIMAX, we replace the missing measurement with the last observed value of the measurements.

A timing diagram showing when variables are logged and when estimates are generated are shown in Fig. 2. We train separate estimators for each estimated variable (NASA-TLX subscale rating in Sia, estimating operator classification accuracy in Sib and Sib'). For each estimator, we perform 4-fold cross-validation on the dataset consisting of the logged data of the 24 participants. We shuffle the data and divide it into 4 folds, each fold consists of 6 participants (25% of the participants). Then we run 4 iterations of training and validation: in each iteration we train the estimator on three folds and validate on the remaining fold. Recall that the data from each participant is a time-series of 25 rounds. In validation, we compute the root mean-squared error (RMSE) given by

$$\text{RMSE} = \sqrt{\frac{\sum_{k=1}^{25}(y_k - \hat{y}_k)^2}{25}},$$

where y_k and $\hat{y}_k$ are the true and estimated values of the estimated variable at round k. We collect the RMSE for the 6 participants in the testing-fold and do so for the four iterations of cross-validation. Since each participant belongs to one fold, we have 24 RMSE values at the end of this procedure, one for each participant. For each estimator, we collect the summary statistics (median and 25% and 75% quantiles) of RMSE.

We use RMSE as the performance metric because it has the same units as the target variable, which makes it easier to interpret. We use 4-fold cross-validation for its stability and reduced noise sensitivity, limiting the risk of overfitting [35]. We now discuss the details of the different schemes.

A. Estimation schemes S1

We view each scheme S1 (i.e., S1a, S1b and S1b') as a supervised learning problem and consider three popular regression models as estimators: i) random forest [36], i) XGBoost [37] and iii) support vector regressors [38], where we train the regression models with the features (i.e., the input variables or attributes of the data) and target values (i.e., the output variable the model aims to estimate based on the input features) described in Table II.

We use the standard implementation of these algorithms from the scikit-learn python library [39]. Random forests and

XGBoost are ensemble learning methods and we use 100 and 1000 estimators, respectively. For support vector regressors we use rbf (radial basis functions) as the kernel parameter from the scikit-learn library.

B. Estimation schemes S2

Schemes S2 (i.e., S2a, S2b and S2b') estimate MW or operator classification accuracy based on a history of observations. For this, we use the dynamic models described below.

1) Supervised learning with a finite window of past observations: The features and target values for schemes S2 when viewed as a regression problem are shown in Table II. We use the same three regression models as estimators and same parameters as in Section IV-A. For each regression model, we computed the average RMSE across the test set for each window length $w \in \{1, \ldots, 10\}$ and chose the value of w with the lowest average RMSE. Details of the choice of w for each estimator are provided in the supplementary material.

2) ARIMAX: We model estimation of operator subscale ratings of the NASA-TLX or operator classification accuracy as a multivariate time series forecasting problem with exogenous input. For a time series $\{x_k\}_{k\geq 1}$, we define the first-order differencing operator Δ as $\Delta x_k = x_k - x_{k-1}$, for $k \geq 1$. Furthermore for any $d > 1$, we define the d-th order differencing operator Δ^d as $\Delta^d x_k = \Delta^{d-1} x_k - \Delta^{d-1} x_{k-1}$. For ease of notation, we define $\Delta^0 x_k = x_k$. An ARIMAX(p, q, s, d) system is given by

$$\Delta^d y_k = \sum_{i=1}^{p} \alpha_i \Delta^d y_{k-i} + \sum_{j=0}^{q} \beta_j u_{k-j} + \sum_{\ell=1}^{s} \gamma_\ell \varepsilon_{k-\ell}, + \varepsilon_k + c \tag{1}$$

where $\{u_k\}_{k\geq 1}$ is the input sequence, $\{y_k\}_{k\geq 1}$ is the output sequence, $\{\varepsilon_k\}_{k\geq 1}$ is the error sequence, (p, q, s) is the model order, d is the order of the differencing operator, and the offset c and the coefficients $\{\alpha_i\}_{i=1}^{p}$, $\{\beta_j\}_{j=0}^{q}$, $\{\gamma_\ell\}_{\ell=1}^{s}$ are real-valued parameters to be learnt.

Because each input and output shown in Table II takes values within a finite interval, say $[V_{\min}, V_{\max}]$, we first transform these values using a non-linear function $\phi : [V_{\min}, V_{\max}] \to (-\infty, \infty)$. To simplify the notation, we use the same symbol ϕ for all these transformations even though the range is different for each variable. The functions ϕ and the ranges $[V_{\min}, V_{\max}]$ for the different variables are described in the supplementary material. Then, for estimating the NASA-TLX subscale ratings and operator classification accuracy, we consider the following inputs and outputs for the model (1):

- $y_k = \phi(\mathrm{RES}_k)$ and $u_k = \mathrm{vec}(\phi(k), \phi(N_k), \phi(\tau_k), \phi(\mathrm{ACC}_k))$ when estimating each NASA-TLX subscale ratings with scheme S2a, where the operator vec stacks all its components into a single vector. Since past values of response to operator ratings are not available, we set $p = d = 0$.
- $y_k = \phi(\mathrm{ACC}_k)$ and $u_k = \mathrm{vec}\,(\phi(k), \phi(N_k), \phi(\tau_k))$ when estimating operator classification accuracy with schemes S2b and S2b'. For S2b', we do not have access to past values of primary OPM, so we set $p = d = 0$.

We use the Python library pmdarima [40] to select the system order and the differencing order (in scheme S2b) and to learn the system coefficients. We also use the library to make estimations based on the inputs and the learned estimator.

3) RNN-LSTM: We consider a Recurrent Neural Network with LSTM gates (RNN-LSTM) to estimate NASA-TLX subscale ratings in S2a and operator classification accuracy in S2b and S2b'. The set of features and target values used for RNN-LSTM are shown in Table II.

An LSTM consists of memory cells that allows the network to learn patterns over long sequences [26]. We use u_k to denote the input and y_k to denote the output of the LSTM. We choose

- $u_k = \mathrm{vec}\,(k, N_k, \mathrm{ACC}_k, \tau_k)$ when estimating each NASA-TLX subscale ratings with scheme S2a,
- $u_k = \mathrm{vec}\,(k, N_k, \mathrm{ACC}_{k-1}, \tau_k)$ when estimating operator classification accuracy with scheme S2b.
- $u_k = \mathrm{vec}\,(k, N_k, \tau_k)$ when estimating operator classification accuracy with scheme S2b'.

Our LSTM network consists of a single layer (as determined through hyperparameter search) with ReLU activation function. The network parameters are optimized using the ADAM optimizer [41]. The number of hidden states for each model associated with NASA-TLX subscale ratings and operator classification accuracy are chosen based on hyperparameter search (performed separately for each subscale and each scheme) and final chosen values are shown in the supplementary material. Estimation with the trained RNN-LSTM network is performed using the forward model given by

$$\hat{y}_k = \sigma(W_{hy} \cdot h_k + b_y),$$

where $\hat{y}_k$ represents estimated operator classification accuracy or NASA-TLX subscale ratings, σ is the activation function (ReLU in our case), W_{hy}, b_y are the learned network parameters and h_k is the hidden state.

4) Kalman filtering on transformed data: We model the quantities to estimate as components of the output of a linear state-space dynamical system, where the inputs are the current round k and the operator taskload N_k and the outputs are average classification time τ_k, operator classification accuracy ACC_k, and operator subscale ratings RES_k.

We apply the same non-linear transformations $\phi\colon [\underline{v}, \overline{v}] \to (-\infty, \infty)$ as for ARIMAX to transform the inputs and outputs of the linear model, and train a linear system of the form:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Gw_k, \\ y_k &= Cx_k + Du_k + w_k, \end{aligned}$$

with state $x_k \in \mathbb{R}^d$, input $u_k = \mathrm{vec}(\phi(N_k), \phi(k))$, and

- $y_k = \mathrm{vec}(\phi(\tau_k), \phi(\mathrm{ACC}_k), \phi(\mathrm{RES}_k)$ when estimating each NASA-TLX subscale ratings with scheme S2a;
- $y_k = \mathrm{vec}(\phi(\tau_k), \phi(\mathrm{ACC}_k))$ when estimating operator classification accuracy with schemes S2b and S2b'.

The disturbance w_k is assumed to be i.i.d. with zero mean and covariance matrix Σ. The system matrices A, B, C, D, G and covariance Σ are real-valued matrices of appropriate dimensions, which are separately estimated for each subscale rating using the `n4sid` algorithm [42] as implemented in the Matlab system identification toolbox. The system order d is

TABLE II
INFORMATION USED FOR THE DIFFERENT ESTIMATION SCHEMES. WE USE THE FOLLOWING SHORT FORMS: FW FOR FINITE WINDOW, RNN FOR RNN-LSTM, HMM FOR IO-HMM, AND KF FOR KALMAN FILTER.FOR ARIMAX AND KF, A NON-LINEAR TRANSFORMATION (NOT SHOWN IN THE TABLE FOR SIMPLICITY) IS PERFORMED BEFORE FITTING THE ESTIMATOR.

Estimator	Variant a (estimate NASA-TLX subscale rating)	Variant b (estimate operator accuracy)	Variant b' (estimate operator accuracy)
		Scheme S1	
	$[k, N_k, \tau_k, \mathrm{ACC}_k]$	$[k, N_k, \tau_k, \mathrm{ACC}_{k-1}]$	$[k, N_k, \tau_k]$
		Scheme S2	
FW, ARIMAX	$[k, N_{k-w:k}, \tau_{k-w:k}, \mathrm{ACC}_{k-w:k}]$	$[k, N_{k-w:k}, \tau_{k-w:k}, \mathrm{ACC}_{k-w:k-1}]$	$[k, N_{k-w:k}, \tau_{k-w:k}]$
RNN, HMM, KF	$[k, N_{1:k}, \tau_{1:k}, \mathrm{ACC}_{1:k}]$	$[k, N_{1:k}, \tau_{1:k}, \mathrm{ACC}_{1:k-1}]$	$[k, N_{1:k}, \tau_{1:k}]$
		Scheme S3	
FW	$[k, N_{k-w:k}, \tau_{k-w:k}, \mathrm{ACC}_{k-w:k}, \mathrm{RES}_{k-1}]$	$[k, N_{k-w:k}, \tau_{k-w:k}, \mathrm{ACC}_{k-w:k-1}, \mathrm{RES}^{\mathrm{all}}_{k-1}]$	$[k, N_{k-w:k}, \tau_{k-w:k}, \mathrm{RES}^{\mathrm{all}}_{k-1}]$
ARIMAX	Not used	$[k, N_{k-w:k}, \tau_{k-w:k}, \mathrm{ACC}_{k-w:k-1}, \mathrm{RES}^{\mathrm{all}}_{k-1}]$	$[k, N_{k-w:k}, \tau_{k-w:k}, \mathrm{RES}^{\mathrm{all}}_{k-1}]$
		Scheme S4	
FW, ARIMAX	$[k, N_{k-w:k}, \tau_{k-w:k}, \mathrm{ACC}_{k-w:k}, \mathrm{RES}_{k-w:k-1}]$	$[k, N_{k-w:k}, \tau_{k-w:k}, \mathrm{ACC}_{k-w:k-1}, \mathrm{RES}^{\mathrm{all}}_{k-w:k-1}]$	$[k, N_{k-w:k}, \tau_{k-w:k}, \mathrm{RES}^{\mathrm{all}}_{k-w:k-1}]$
RNN, HMM, KF	$[k, N_{1:k}, \tau_{1:k}, \mathrm{ACC}_{1:k}, \mathrm{RES}_{1:k-1}]$	$[k, N_{1:k}, \tau_{1:k}, \mathrm{ACC}_{1:k-1}, \mathrm{RES}^{\mathrm{all}}_{1:k-1}]$	$[k, N_{1:k}, \tau_{1:k}, \mathrm{RES}^{\mathrm{all}}_{1:k-1}]$

automatically chosen by the algorithm based on Hankel singular value decomposition. The system orders for the different estimators are presented in the supplementary material.

Let z_k denote the variable that we are interested in estimating, i.e., $z_k = \phi(\mathrm{RES}_k)$ in S2a and $z_k = \phi(\mathrm{ACC}_k)$ in S2b and S2b'. In both cases, z_k is an element of y_k; let C_z and D_z denote the corresponding rows of C and D. Only a subset of the data is available for estimation, as shown in Table II. For a generic scheme, we denote this data by $\mathrm{data}(k)$. The estimate $\hat{z}_k$ of z_k is generated as

$$\hat{z}_k = C_z\, \mathbb{E}[x_k \mid \mathrm{data}(k)] + D_z u_k,$$

where the state estimate $\mathbb{E}[x_k|\mathrm{data}(k)]$ is recursively updated using Kalman filtering [43].

5) Non-linear filtering based on IO-HMMs: We model the quantities to estimate as components of the output of an IO-HMM [44], where the inputs are the current round k and the operator's taskload N_k and the outputs are average classification time τ_k, (quantized) operator classification accuracy $\lfloor \mathrm{ACC}_k \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function, and each NASA-TLX subscale ratings RES_k. We assume that the IO-HMM has a hidden discrete state $x_k \in \{1, \dots H\}$, where the integer H is the size of the state space, and consider the following two transition kernels P and Q:

$$P(j \mid i, u) = \Pr(x_{k+1} = j \mid x_k = i, u_k = u),$$
$$Q(y \mid i, u) = \Pr(y_k = y \mid x_k = i, u_k = u),$$

where $u_k = \mathrm{vec}(N_k, k)$ is the input and

- y_k=vec($\tau_k, \lfloor \mathrm{ACC}_k \rfloor, \mathrm{RES}_k$) when estimating each NASA-TLX subscale ratings with scheme S2a;
- $y_k = \mathrm{vec}(\tau_k, \lfloor \mathrm{ACC}_k \rfloor)$ when estimating operator classification accuracy with schemes S2b and S2b'.

The kernels P and Q are learned using an extended expectation maximization (EM) algorithm [45]. We performed a hyper-parameter search with different values of H, and chose the value that has the lowest average RMSE score averaged over the test experiment set. The number of hidden states chosen for the estimation of operator classification accuracy and NASA-TLX subscale ratings are provided in the supplementary material.

As in the Kalman filter setting, let z_k denote the variable that we are interested in estimating, which is a component of y_k. Let Q_z denote the corresponding marginalization of the kernel Q, i.e.,

$$Q_z(z|i, u) = \Pr(z_k = z|x_k = i, u_k = u).$$

Moreover, as in the Kalman filtering case, let $\mathrm{data}(k)$ denote the data available at the time of estimation, as shown in Table II. We then recursively update the belief

$$\alpha_k(i) \triangleq \Pr(x_k = i \mid \mathrm{data}(k))$$

using the forward algorithm of [46], which is implemented in Python. We then compute the MMSE (minimum mean squared error) estimate of z_k as

$$\hat{z}_k = \sum_z z \sum_{i=1}^{H} \alpha_k(i) Q_z(z \mid i, u_k). \tag{2}$$

C. Estimation schemes S3

Schemes S3 (i.e., S3a, S3b and S3b') use the history of both primary and secondary OPM measurements and task characteristics along with the most recent value of NASA-TLX subscale

ratings. When estimating operator classification accuracy, we include as inputs all NASA-TLX subscale ratings, denoted by $\mathrm{RES}_k^{\mathrm{all}}$. On the other hand, to estimate a given subscale rating RES_k we include as input this particular subscale rating only. We consider the following schemes.

1) Supervised learning with a finite window of past observations: We follow the framework of Section IV-B1, with the features and target values shown in Table II.

2) ARIMAX: We use ARIMAX only to estimate operator classification accuracy (schemes S3b and S3b'), following the framework of Section IV-B2. The inputs are $u_k = \mathrm{vec}(\phi(k), \phi(N_k), \phi(\tau_k), \phi(\mathrm{RES}_{k-1}^{\mathrm{all}}))$ and the outputs are $y_k = \phi(\mathrm{ACC}_k)$. For S3b', we set $p = d = 0$ to exclude past values of primary OPM measurements.

D. Estimation schemes S4

Compared to schemes S3, schemes S4 (i.e., S4a, S4b and S4b') add the history of NASA-TLX subscale ratings when estimating both operator classification accuracy and next operator subscale ratings.

1) Supervised learning with a finite window of past observations and RNN-LSTM: For these two schemes, we follow the frameworks of Section IV-B1 and IV-B3 respectively, with the features and target values shown in Table II.

2) ARIMAX: We follow the framework of Section IV-B2, where we use the following inputs and outputs:

- $u_k = \mathrm{vec}(\phi(k), \phi(N_k), \phi(\tau_k), \phi(\mathrm{ACC}_k))$ and $y_k = \phi(\mathrm{RES}_k)$ when estimating each NASA-TLX subscale ratings with scheme S4a;
- $u_k = \mathrm{vec}(\phi(k), \phi(N_k), \phi(\tau_k), \phi(\mathrm{RES}_{k-1}^{\mathrm{all}}))$ and $y_k = \phi(\mathrm{ACC}_k)$ when estimating operator classification accuracy with schemes S4b and S4b'. For scheme S4b', we set $p = d = 0$ to exclude past primary OPM measurements.

3) Kalman Filtering on transformed data: We use the same modeling framework as presented in Section IV-B4, and learn linear models with inputs $u_k = \mathrm{vec}\,(\phi(k), \phi(N_k))$ and outputs

- $y_k = \mathrm{vec}\,(\phi(\tau_k), \phi(\mathrm{ACC}_k), \phi(\mathrm{RES}_k))$ when estimating each NASA-TLX subscale ratings with scheme S4a;
- $y_k = \mathrm{vec}\,(\phi(\tau_k), \phi(\mathrm{ACC}_k), \phi(\mathrm{RES}_k^{\mathrm{all}}))$, when estimating operator classification accuracy (S4b and S4b').

The estimation procedure is then exactly the same as in Section IV-B4 where $\mathrm{data}(k)$ used for estimation is as specified in Table II.

4) Non-linear filtering based on IO-HMMs: We use a modeling framework for IO-HMM similar to the one presented in Section IV-B and consider the following as the input and output of the IO-HMM:

- $u_k = \mathrm{vec}\,(N_k, k)$
- $y_k = \mathrm{vec}\,(\tau_k, \lfloor \mathrm{ACC}_k \rfloor, \mathrm{RES}_k)$ when estimating each NASA-TLX subscale rating with scheme S4a.
- $y_k = \mathrm{vec}\,(\tau_k, \mathrm{RES}_k^{\mathrm{all}}, \lfloor \mathrm{ACC}_k \rfloor)$ when estimating operator classification accuracy with schemes S4b and S4b'.

The estimation procedure is then exactly the same as in Section IV-B5 where $\mathrm{data}(k)$ used for estimation is as specified in Table II.

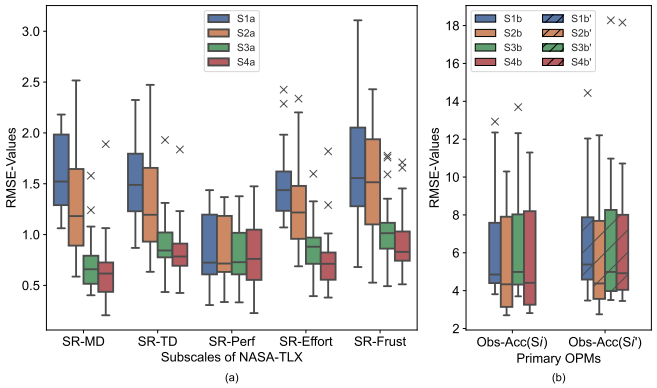

Fig. 3. Box plot comparing RMSE for the best estimators in schemes S1-S4 for (a) NASA-TLX subscale ratings and (b) operator classification accuracy.

E. Estimation Schemes S5

Scheme S5 (i.e., S5a, S5b, and S5b') is similar to S4 except that the NASA-TLX subscale ratings are available every t_m rounds, where t_m is called the measurement interval. We consider the same estimators as in S4. For finite window regressors, ARIMAX, and RNN-LSTM we carry forward the last observed values for the missing measurements during both training and estimation. For Kalman filtering and IO-HMMs, we use time-varying observation channels, i.e., the matrix C in the case of Kalman filtering and the kernel Q in the case of IO-HMMs change at different rounds k (for both training and estimation) to capture the fact that RES_k is available every t_m interval.

V. Results

We use the following abbreviations to present our results. SR-MD: Self reported mental demand, SR-TD: Self reported temporal demand, SR-Perf: Self reported performance, SR-Effort: Self reported effort, SR-Frust: Self reported frustration and Obs-Acc: Operator classification accuracy.

For each scheme, the summary statistics of each estimator's RMSE are given in the supplementary material, where we show the median and quantiles in the form $median[q_1, q_3]$ where q_1 and q_3 represents 25% and 75% quantiles of the data respectively. For scheme S5, we only show the data for three values of measurement interval: $t_m \in \{3, 5, 8\}$. To compare across schemes, we pick the estimator with the smallest median as the representative estimator for each scheme. The performance values of the representative estimators for schemes S1–S4 are shown in Figure 3. Those for scheme S5 for different values of measurement interval t_m are shown in Figure 4, with the performance of S4 also shown for comparison.

Figure 3(a) suggests that for estimating the NASA-TLX subscale ratings in general the median RMSE of the different schemes are in the order S4a $\leq$ S3a $\leq$ S2a $\leq$ S1a, [2] except for the performance subscale, where all schemes perform roughly the same. Figure 3(b) shows the box plot of RMSE values

[2]The notation Sia $\leq$ Sja means that scheme Sia performs better than scheme Sja.

TABLE III
p-VALUES FOR T-TESTS WHEN COMPARING BETWEEN SCHEMES S1-S4. p-VALUES BELOW 0.1 ARE SHOWN IN BOLD, THOSE BELOW 0.01 ARE ALSO UNDERLINED.

Scheme comparison	SR-MD	SR-TD	SR-Perf	SR-Effort	SR-Frust	Obs-Acc
S2 < S1	**0.038**	**0.075**	0.491	**0.083**	0.314	0.459
S3 < S2	$\underline{\mathbf{4.53 \times 10^{-6}}}$	$\underline{\mathbf{1.4 \times 10^{-3}}}$	0.425	$\underline{\mathbf{8.58 \times 10^{-5}}}$	$\underline{\mathbf{8.6. \times 10^{-4}}}$	0.305
S4 < S3	0.21	0.17	0.35	**0.091**	**0.015**	0.78

when operator classification accuracy is estimated using Sib and Sib' respectively where $i \in \{1, 2, 3, 4\}$. Figure 3(b) suggests that for estimating operator classification accuracy, the median RMSE values for all schemes S1b–S4b and S1b'–S4b' are roughly the same.

In the following, we answer the research questions posed in Section II using the data shown on Figure 3 and Figure 4. In addition, we perform a series of independent samples t-tests [47]. Each test compares two schemes. We use the notation Si < Sj to denote the t-test where the null hypothesis is that the average RMSE of scheme Si is equal to that of scheme Sj while the alternative hypothesis is that the average RMSE of scheme Si is less than that of scheme Sj. The p-values for the t-test when comparing between schemes S1–S4 are shown in Table III, where the more statistically significant test results are outlined.

Similarly, when comparing between schemes Si and the corresponding Si' we use the notation Si < Si' to denote the t-test where the null hypothesis is that the average RMSE of Si is equal to Si' while the alternative hypothesis is that the average RMSE of Si is less than Si'.

A. Answer to Q1

To answer Q1, we test S2 < S1. Figure 3 indicates that incorporating the history of passive measurements and task characteristics helps reduce the estimation error of SR-MD, SR-TD and SR-Effort, with stronger statistical evidence that this is the case for SR-MD (p-value 0.038). However, there is insufficient evidence to support this conclusion for the SR-Perf and SR-Frust NASA-TLX subscales or for operator classification accuracy.

B. Answer to Q2

Question Q2 is answered by two tests: S3 < S2, for which we have strong statistical evidence in the case of SR-MD, SR-TD, SR-Effort, and SR-Frust; and S4 < S3, for which we have statistical evidence for SR-Frust and, to a lesser extent, for SR-Effort. Thus, we conclude that including the *last* NASA-TLX subscale rating reduces the estimation error of these ratings for mental demand, temporal demand, effort and frustration. Moreover, including additional historical measurements of NASA-TLX subscale ratings seems to help reduce the estimation error of the effort and frustration subscale ratings. However, there is insufficient evidence to support the same conclusions for other NASA-TLX subscale ratings or for operator classification accuracy.

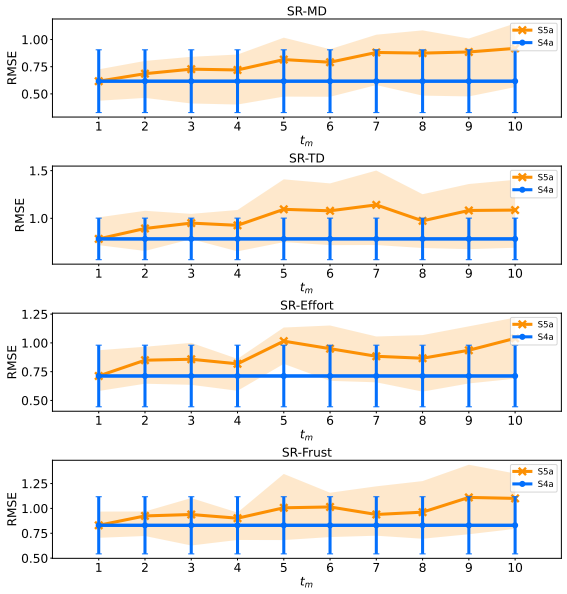

Fig. 4. Comparison of RMSE of schemes S4a and S5a. The median is shown by solid lines and the interquantile range is shown by vertical error bars for S4a and by a shaded region for S5a.

C. Answer to Q3

As reported in the last section, for Q2 we found insufficient evidence to conclude that past measurements of NASA-TLX subscale ratings help in estimating SR-Perf or Obs-Acc. So, for Q3, we focus on the remaining NASA-TLX subscales: SR-MD, SR-TD, SR-Effort, and SR-Frust. The RMSE of different estimators for S5a for different values of the measurement interval t_m are shown in Fig. 4, where the RMSE for S4a is also shown for comparison.

Figure 4 highlights the trade-off between the measurement interval and RMSE. A smaller measurement interval implies a smaller RMSE, but at the expense of the operator having to answer the questionnaires more frequently, which could distract from the primary task. A larger measurement interval reduces the frequency of intrusive questionnaires but at the expense of the larger RMSE. Figure 4 shows the exact trade-off between performance and measurement interval, which can be useful for choosing the value of t_m in specific applications. Given that each round in this experiment lasted two minutes, the figure also sheds some light on the dynamics of MW (as measured by the NASA-TLX) and the extent of its temporal correlations, which can be exploited by dynamic estimators.

D. Answer to Q4

Q4 is answered by a series of tests Sib < Sib' for $i \in \{1, 2, 3, 4\}$. None of the tests are statistically significant (p values are between 0.39 to 0.77, see supplementary material). Therefore, we do not have evidence to conclude that past observations of primary OPM measurements reduce the RMSE (estimation error) for operator classification accuracy in this experiment, once the secondary OPM measurements and task characteristics are taken into account.

VI. Discussion

Our analysis shows that incorporating the last observed NASA-TLX subscale ratings, the history of primary and secondary OPM measurements and the task characteristics improves the estimation of all NASA-TLX subscale ratings, except for self-reported performance. Moreover, adding more history of past NASA-TLX subscale ratings further helps estimating the MW components captured by the effort and frustration subscales of the NASA-TLX. Therefore, we have evidence to support the dynamic nature of several components of MW, and the usefulness of employing dynamic models for their estimation.

On the other hand, we cannot conclude for our experiment that past measurements of subscale ratings, primary and secondary OPMs and task characteristics improve the estimation of self-reported performance in the NASA-TLX questionnaires, or of the actual operator classification accuracy. This may be due to insufficient data, or more plausibly due to specific features of our experiment. In separate analysis, we have also tried to estimate operator accuracy at round k using only the NASA-TLX subscale ratings collected at the end of that round, but the estimation performance was worse than by simply using the secondary OPM measurements (average classification time) for the round, which moreover are obtained more easily and earlier than subjective ratings. Similar dissociation between MW and task performance has been previously reported in the literature [48]–[51]. This motivates the use of additional more objective measurements to predict operator performance, such as secondary OPMs or physiological measurements. In particular, for our experiment a simple regression model as in Scheme S1b' based on secondary OPMs and task characteristics was sufficient to estimate operator accuracy. In any case, further work is needed to identify conditions under which MW or other cognitive state assessments can help predict future operator performance, given the interest in using such assessments to adjust the level of automation dynamically.

Based on our analysis, we can make prescriptive recommendations for choosing schemes when estimating an operator's MW (or, equivalently here, NASA-TLX subscale ratings). In certain applications, obtaining direct subjective measurements of MW is difficult because administering a questionnaire can interfere with the operator's task. In those cases, estimation of MW can be performed by monitoring operator performance, as in schemes S1a and S2a. In our experiment, we found evidence that using dynamic estimators leveraging past performance measurements can help improve MW estimates along several dimensions. In applications where subjective measurements can be obtained, we obtained strong evidence that using the most recent of these measurements (in our experiment, obtained two minutes prior at the end of the previous round) is helpful to predict most dimensions of MW. However, the benefits of leveraging additional past subjective measurements were more limited. Furthermore, we evaluated through scheme 5a that the MW estimation performance steadily degrades, within minutes, as the latest subjective measurements get older. This sheds some light on the temporal dynamics of the characteristics MW captured by the NASA-TLX questionnaire.

VII. Conclusion

In this paper, we studied dynamic estimation of operator mental workload (MW), as measured by different subscales of the NASA-TLX questionnaire, and operator classification accuracy, using various estimation schemes. In our analysis, using dynamic estimation led to significant reduction in estimation error for most NASA-TLX subscales; however, the reduction in estimation error was limited for the NASA-TLX self-reported performance subscale and for actual operator classification accuracy, once the information contained in secondary performance metrics and task characteristics was taken into account.

In our current setup, the workload at each round was fixed before the start of the experiment. An interesting and important future direction is to evaluate dynamic adaptation of task load based on estimated values of MW and operator performance measurements.

References

[1] P. A. Hancock and G. Matthews, "Workload and performance: Associations, insensitivities, and dissociations," *Human factors*, vol. 61, no. 3, pp. 374–392, 2019.

[2] C. D. Wickens, S. E. Gordon, Y. Liu, and J. Lee, *An introduction to human factors engineering*. Pearson Prentice Hall Upper Saddle River, NJ, 2004, vol. 2.

[3] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs," *J. Cogn. Engineering and Decision Making*, vol. 2, no. 2, pp. 140–160, 2008.

[4] K. M. Feigh, M. C. Dorneich, and C. C. Hayes, "Toward a characterization of adaptive systems: A framework for researchers and system designers," *Human factors*, vol. 54, no. 6, pp. 1008–1024, 2012.

[5] S. Rubio, E. Díaz, J. Martín, and J. M. Puente, "Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods," *Appl. Psychol.*, vol. 53, no. 1, pp. 61–86, 2004.

[6] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Adv. Psych.* Elsevier, 1988, vol. 52, pp. 139–183.

[7] A. Marinescu, S. Sharples, A. Ritchie, T. S. López, M. McDowell, and H. Morvan, "Exploring the relationship between mental workload, variation in performance and physiological parameters," *IFAC-PapersOnLine*, vol. 49, no. 19, pp. 591–596, 2016.

[8] H. Ayaz and F. Dehais, *Neuroergonomics: The brain at work and in everyday life*. Academic Press, 2018.

[9] C. D. Wickens, "Multiple resources and performance prediction," *Theoretical issues in ergonomics science*, vol. 3, no. 2, pp. 159–177, 2002.

[10] ——, "Processing resources and attention," in *Multiple task performance*. CRC Press, 2020, pp. 3–34.

[11] G. F. Wilson and C. A. Russell, "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," *Human factors*, vol. 45, no. 4, pp. 635–644, 2003.

[12] Z. Yin, M. Zhao, W. Zhang, Y. Wang, Y. Wang, and J. Zhang, "Physiological-signal-based mental workload estimation via transfer dynamical autoencoders in a deep learning framework," *Neurocomputing*, vol. 347, pp. 212–229, 2019.

[13] J. Zhang and S. Li, "A deep learning scheme for mental workload classification based on restricted boltzmann machines," *Cogn., Tech. & Work*, vol. 19, pp. 607–631, 2017.

[14] Ş. H. Aksu, E. Çakıt, and M. Dağdeviren, "Mental workload assessment using machine learning techniques based on eeg and eye tracking data," *Applied Sciences*, vol. 14, no. 6, p. 2282, 2024.

[15] C. J. Lin and R. P. Lukodono, "Classification of mental workload in human-robot collaboration using machine learning based on physiological feedback," *Journal of Manufacturing Systems*, vol. 65, pp. 673–685, 2022.

[16] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," *Appl. Ergonomics*, vol. 74, pp. 221–232, 2019.

[17] J. R. Kintz, N. T. Banerjee, J. Y. Zhang, A. P. Anderson, and T. K. Clark, "Estimation of subjectively reported trust, mental workload, and situation awareness using unobtrusive measures," *Human Factors*, p. 00187208221129371, 2022.

[18] P. Wang, W. Fang, and B. Guo, "A measure of mental workload during multitasking: Using performance-based timed petri nets," *Intl. Journal of Industrial Ergonomics*, vol. 75, p. 102877, 2020.

[19] K. Akash, K. Polson, T. Reid, and N. Jain, "Improving human-machine collaboration through transparency-based feedback–part I: Human trust and workload model," *IFAC-PapersOnLine*, vol. 51, no. 34, pp. 315–321, 2019.

[20] M. A. Shahab, M. U. Iqbal, B. Srinivasan, and R. Srinivasan, "HMM-based models of control room operator's cognition during process abnormalities. 1. formalism and model identification," *J. Loss Prev. Process Ind.*, vol. 76, p. 104748, 2022.

[21] X. Fan, P.-C. Chen, and J. Yen, "Learning HMM-based cognitive load models for supporting human-agent teamwork," *Cogn. Syst. Res.*, vol. 11, no. 1, pp. 108–119, 2010.

[22] B. Cain, "A review of the mental workload literature," *DTIC Document*, 2007.

[23] C. D. Wickens, W. S. Helton, J. G. Hollands, and S. Banbury, *Engineering psychology and human performance*. Routledge, 2021.

[24] D. Kratchounova, I. Choi, T. C. Mofle, L. Miller, S. Stevenson, and M. Humphreys, "Exploring the relationship between flight technical error and NASA-TLX subscale ratings when using HUD localizer takeoff guidance in lieu of currently required infrastructure," FAA, Tech. Rep. DOT/FAA/AM-21/29, 2021.

[25] J. Heard, P. Baskaran, and J. A. Adams, "Predicting task performance for intelligent human-machine interactions," *Frontiers in Neurorobotics*, vol. 16, p. 973967, 2022.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] J. M. Shine, P. G. Bissett, P. T. Bell, O. Koyejo, J. H. Balsters, K. J. Gorgolewski, C. A. Moodie, and R. A. Poldrack, "The dynamics of functional brain networks: integrated network states during cognitive task performance," *Neuron*, vol. 92, no. 2, pp. 544–554, 2016.

[28] L. M. Ma, M. Ijtsma, K. M. Feigh, and A. R. Pritchett, "Metrics for human-robot team design: A teamwork perspective on evaluation of human-robot teams," *ACM Trans. Hum. Robot Interact. (THRI)*, vol. 11, no. 3, pp. 1–36, 2022.

[29] C. Wu, Y. Liu, and B. Lin, "A queueing model based intelligent human–machine task allocator," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1125–1137, 2012.

[30] K. Savla and E. Frazzoli, "A dynamical queue approach to intelligent task management for human operators," *Proc. of the IEEE*, vol. 100, no. 3, pp. 672–686, 2011.

[31] A. Cao, K. K. Chintamani, A. K. Pandya, and R. D. Ellis, "Nasa tlx: Software for assessing subjective mental workload," *Behavior research methods*, vol. 41, pp. 113–117, 2009.

[32] M. L. Bolton, E. Biltekoff, and L. Humphrey, "The mathematical meaninglessness of the NASA task load index: A level of measurement analysis," *IEEE Trans. Human-Mach. Syst.*, 2023.

[33] F. Vachon, D. Lafond, B. R. Vallieres, R. Rousseau, and S. Tremblay, "Supporting situation awareness: A tradeoff between benefits and overhead," in *Intl. multi-disciplinary Conf. cognitive methods in situation awareness and decision support*. IEEE, 2011, pp. 284–291.

[34] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human–autonomy teaming: A review and analysis of the empirical literature," *Human factors*, vol. 64, no. 5, pp. 904–938, 2022.

[35] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[36] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[37] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Sigkdd Intl. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[38] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[40] T. G. Smith *et al.*, "pmdarima: Arima estimators for Python," 2017–, [Online; accessed 2024-06-06]. [Online]. Available: http://www.alkaline-ml.com/pmdarima

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[42] P. Van Overschee and B. De Moor, "N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems," *Automatica*, vol. 30, no. 1, pp. 75–93, 1994.

[43] B. Anderson and J. B. Moore, "Optimal filtering," *Prentice-Hall Information and System Sciences Series*, 1979.

[44] Y. Bengio and P. Frasconi, "An input output HMM architecture," *Adv. Neural Inf. Process. Syst.*, vol. 7, 1994.

[45] ——, "Input-output hmms for sequence processing," *IEEE Trans. Neural Netw.*, vol. 7, no. 5, pp. 1231–1249, 1996.

[46] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *Ann. Math. Stat.*, vol. 37, no. 6, pp. 1554–1563, 1966.

[47] D. A. Freedman, *Statistical models: theory and practice*. Cambridge University Press, 2009.

[48] M. Vidulich and C. Wickens, "Causes of dissociation between subjective workload measures and performance: Caveats for the use of subjective assessments," *Appl. Ergonomics*, vol. 17, no. 4, pp. 291–296, 1986. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0003687086901328

[49] M. Myrtek, E. Deutschmann-Janicke, H. Strohmaier, W. Zimmermann, S. Lawerenz, G. Brügner, and W. Müeller, "Physical, mental, emotional, and subjective workload components in train drivers," *Ergonomics*, vol. 37, no. 7, pp. 1195–1203, 1994.

[50] D. De Waard, "The measurement of drivers' mental workload," Ph.D. dissertation, University of Groningen, 1996.

[51] M. S. Young, K. A. Brookhuis, C. D. Wickens, and P. A. Hancock, "State of science: mental workload in ergonomics," *Ergonomics*, vol. 58, no. 1, pp. 1–17, 2015.

PLACE PHOTO HERE

Raihan Seraj (Student Member, IEEE) is a PhD student at the Department of Electrical and Computer Engineering at McGill University, Canada. He received his BSc. in Electrical and Electronics Engineering from Islamic University of Technology, Dhaka, Bangladesh in 2015 and his MEng in Electrical Engineering at McGill University, Montreal, Canada in 2019. His research interests include reinforcement learning, human in the loop systems and time series modelling.

PLACE PHOTO HERE

Aditya Mahajan (Senior Member, IEEE) is Professor of Electrical and Computer Engineering at McGill University, Montreal, Canada. He received the B.Tech degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, India in 2003 and the MS and PhD degrees in Electrical Engineering and Computer Science from the University of Michigan, Ann Arbor, USA in 2006 and 2008, respectively. He serves or has served as Associate Editor of Transactions on Automatic Control, Control Systems Letters, and Math. of Control, Signal, and Systems.

PLACE PHOTO HERE

Jerome Le Ny (Senior Member, IEEE) is a Professor of Electrical Engineering at Polytechnique Montreal. He received the engineering degree from the École Polytechnique, France, in 2001, the M.Sc. degree in Electrical Engineering from the University of Michigan, Ann Arbor, MI, USA, in 2003, and the Ph.D. degree in Aeronautics and Astronautics from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. His research interests include stochastic control, navigation systems, networked and multiagent control systems, with applications to autonomous robots and intelligent infrastructure systems. He currently serves as an Associate Editor of the IEEE Transactions on Robotics.