A Human Workload Assessment Algorithm for Collaborative Human-Machine Teams

Jamison Heard[1], Caroline E. Harriott[2], and Julie A. Adams, Senior Member[3]

Abstract— **Mass casualty events caused by a biological weapon require fully capable first response teams. However, human first responders are equipped with protective gear, which limits their capabilities to complete tasks. Robots can be employed to work collaboratively with the first responders in order to augment the human's reduced abilities. The robot needs to understand and adapt to the human's workload level in order for the human-machine team to effectively complete tasks. The automatic detection of human workload levels can provide valuable insight into the human's capabilities, as workload has a direct relationship with task performance. The robot can monitor the objective metrics of the human's workload level in order to accurately estimate workload via a workload assessment algorithm. The algorithm must be able to assess overall workload and the components of workload, in order for the robot to correctly adapt its interactions or reallocate tasks among the team. A novel workload assessment algorithm that provides an accurate estimate of overall workload and each workload component is presented and evaluated. The algorithm is capable of distinguishing between high and low workload conditions; however, the algorithm's workload values correlate poorly to a generated workload model. Modifications to enhance the algorithm's capabilities are discussed and will be investigated in future work.**

I. Introduction

Human search and rescue teams are employed in first response environments, such as a mass casualty event. However, the human's ablities are severly reduced, due to wearing personal protective gear. A robot can be teamed with the human to augment the human's reduced capabilities and collaboratively complete tasks [1], [2]. Such a collaboration requires the robot to understand and adapt its interactions or reallocate tasks based on the human's workload level, as a too high (overload) or too low (underload) workload level can negatively impact task performance [3], [4].

It is important that the robot understand the human's overall workload state, as well as the distinct components contributing to the workload state. Overall workload can be decomposed into five components: cognitive, auditory, speech, visual, and physical [5]. Tasks incorporate varying proportions of each component, i.e., triaging a victim may require little physical workload, while searching for and identifying hazards can require medium to high physical workload. Robots can use knowledge of the task and human's workload level to optimally adapt its interactions and allocate tasks among the team members. For example, a human triaging a victim and communicating with incident command may have an overloaded workload state. The robot can detect this state and reallocate the communication task to itself in order to improve the team's overall task performance.

Physiological indicators, such as heart-rate variability, provide insight into the human's workload state [6], [7]. A robot or system can use a workload assessment algorithm to analyze these physiological indicators, in order to produce an accurate real-time workload estimate [8], [9]. Typical workload assessment algorithms incorporate machine-learning techniques to classify cognitive workload, typically only for the overload state. These algorithms fail to estimate the other four workload components; thus, not providing a correct assessment of the overall workload state. A robotic teammate needs an accurate assessment of the entire workload state in order to correctly adapt its interactions or reallocate tasks. There is a need for a workload assessment algorithm that is capable of estimating overall workload and the distinct components contributing to the overall workload state.

A contribution of this paper is a workload assessment algorithm that assesses each workload component and intelligently aggregates the component values into an overall workload value. The algorithm relies on machine-learning and workload models to correctly estimate a human's overall workload state. Section II details the workload assessment algorithm, while Section III provides the experimental design. Section IV analyzes the results and Section V provides an in-depth discussion of the results and details future work to enhance the algorithm's capabilities.

II. System Description

The workload assessment algorithm relies on physiological metrics to provide real-time workload assessment information. The algorithm's incorporated physiological metrics and their associated response to workload is provided in Table I. A grey cell represents a correlation between the metric and the corresponding workload component.

Heart rate (HR), respiration rate (RR), skin temperature (ST), and postural magnitude (PM) physiological metrics were measured via the Biopac BioHarness$^{\text{TM}}$ every four miliseconds [21]. The other physiological metrics are either derived or simulated. Heart-rate variability (HRV) is the average beat-to-beat interval, which is derived from the Electrocardiogram signal collected via the BioHarness$^{\text{TM}}$ [10], [11]. Posture sway (PS) is calculated as the variance

*This work was supported by NASA Cooperative Agreement Number NNX16AB24A

[1]Jamison Heard is a PhD student in the Electrical and Computer Engineering Department, Vanderbilt University, Nashville, TN 37235

[2]Caroline E. Harriott is a researcher scientist at Draper Labs, Cambridge, Massachusetts 02139

[3] Julie A. Adams is a Professor in Computer Science and Robotics at Oregon State University, Corvallis, OR 97330

TABLE I

Correlation of Physiological Workload Metrics to Overall workload and associated components

Metric	Response	Overall	Cognitive	Auditory	Speech	Visual	Physical
Heart Rate Variability (HRV) [10], [11], [6]	decreases	■	■				
Heart Rate (HR) [6], [12]	increases	■					■
Respiration Rate (RR) [13], [14]	decreases	■			■		■
Skin Temperature (ST) [15], [16]	decreases	■	■				■
Noise Level (NL) [17], [18]	increases	■	■	■	■		
Postural Magnitude (PM) [19]	increases	■					■
Posture Sway (PS) [20]	increases	■					■

of the postural magnitude, while the noise level (NL) is a simulated static noise level with additive Gaussian noise.

The physiological data from a prior evaluation [22] were subjected to a 30 second non-overlapping moving average before being filtered. Outliers were replaced with a realistic maximum value. For example, a heart rate of 240 bpm is replaced with 180 bpm, as max heart-rate is determined by $HR_{max} = 208 - 0.7 * age$ [23]. The filtered physiological data were mapped to a value (0-100) using three approaches: normalization, regression, and a neural network. Normalization maps data using: $NewValue = (CurrentValue - min)/(max - min)$, where max and min are a data-set's maximum and minimum value, respectively. Regression finds a line of best fit between the physiological data (i.e., x value) and a desired value (i.e., y value) [24]. Neural networks are intended to mimic computation within the human brain and consist of at least three layers: input, processing, and output, where the processing layer contains adaptive weights that are tuned by a learning algorithm [24].

Normalization is based on either an individual participant's physiological metric's max and min values or the max and min values for the entire participant population. The regression and neural network mappings are trained using 75% of the physiological data and tested using the remaining 25%. Normalized overall subjective ratings collected after each task are used as the desired workload level estimate for the training. Neural Networks are prone to becoming stuck at local optimas; thus, each neural network is trained multiple times, until the mean-squared error no longer significantly decreases. Five, ten, and twenty neurons for the neural network's hidden layer were examined, but no significant difference was found between the respective outputs.

The mapped physiological data are aggregated using uniform weights to generate a workload value for each workload component: cognitive (W_C), physical (W_P), auditory (W_A), speech (W_S), and visual (W_V). The workload component aggregation equations are:

$$W_C = 0.25 * (hr + (100 - hrv) + (100 - st) + nl), \quad (1)$$

$$W_P = 0.20 * (hr + (100 - rr) + (100 - st) + pm + ps), \quad (2)$$

$$W_A = nl, \quad (3)$$

$$W_S = 20 * SubjectiveSpeechWorkload, \quad (4)$$

$$W_V = 20 * SubjectiveVisualWorkload. \quad (5)$$

Respiration rate is the only physiological metric that correlates to the speech workload component, but respiration rate has a low sensitivity to workload variations [7]. Further, no collected physiological metrics correlate to visual workload components. Thus, in situ subjective workload responses were used to represent speech and visual workload. The individual workload component values are aggregated to generate an estimate of the overall workload state:

$$W_O = w_C * W_C + w_A * W_A + w_S * W_S + w_V * W_V + w_P * W_P, \quad (6)$$

where W_O is the overall workload value and w_X ($X = C, A, S, V, P$) represents the respective workload component's task weighting. The task weightings are extrapolated from the task's workload model, which was generated by IMPRINT Pro [25] and are presented in prior work [22].

III. Experimental Design

A. Data Collection from Prior Evaluation

The physiological data was collected during a time-structured human-machine teaming evaluation, which required participants to work with a robot to complete first response tasks. Eighteen participants, teamed with a Pioneer 3 robot previously completed four fifteen minute tasks and were assigned to either a low or high workload condition for each task [22]. The results from ten participants was used to analyze the workload assessment algorithm. Table II provides the workload conditions the ten participants were assigned.

TABLE II

Participant Task Workload Levels

Participant	Task 1	Task 2	Task 3	Task 4
P1	Low	High	Low	High
P2	High	Low	Low	Low
P3	High	High	High	High
P4	High	High	High	Low
P5	Low	Low	High	Low
P6	Low	Low	High	High
P7	High	High	Low	High
P8	Low	High	High	High
P9	Low	High	High	Low
P10	High	Low	High	Low

Each participant completed a demographic questionnaire and received a three minute briefing. Participants also received task specific training prior to each task. The participants' mean age was 22.63 (St. Dev. = 6.16) and each participant slept a median of 6.5 hours the night before.

The evaluation's first task involved participants identifying suspicious items in photographs on a Google Nexus 7 tablet. The participant marked any item in the picture deemed suspicious by circling the item or providing a note on the picture. Folders containing three pictures were sent to the participant at predetermined times, where the robot indicated that a new folder arrived with an audible beep. The low and high workload conditions were determined by the number of folders sent to the participant, two and four, respectively.

The second task required searching an environment for hazardous materials, which the participant had to photograph. The environment consisted of an academic building's hallway, offices, and laboratory classrooms. Participants wore equipment to simulate personal protective equipment, i.e., safety gloves, goggles, a dusk mask, and a ten-pound backpack. The robot scanned the floor for items and collected air samples and audibly beeped when information needed to be sent to incident command at predefined times, i.e., 3:45, 7:30, 11:15, and 15:00. The low workload condition contained four items, while the high condition contained eight items.

The third and fourth tasks were solid and liquid containment sampling tasks, respectively. The participant searched for and collected solid or liquid samples in a room, based on the guidelines provided by the robot. The guidelines were based on government standards for sample collection [26]. The samples were taken from containers holding colored solids and liquids, i.e., colored sand and dyed water. Participants wore protective goggles and gloves to simulate personal protective equipment. The participants wore a weighted backpack in the liquid containment sampling task, in order to increase the amount of physical workload. Workload was determined by the number of samples to be collected, two for the low workload and four for the high workload condition.

Each participant rated six workload channels, (cognitive, motor, tactile, auditory, speech, and visual), from 1 - 5 after each task, where an overall rating is the aggregate of the six channels. Previous analysis showed that there is a significant difference in the overall subjective ratings between the low and high workload conditions [22].

Each task and workload condition was modeled using IMPRINT Pro, to ensure that the correct workload state was ellicited [22]. The tasks were broken down into subtasks, i.e., opening a container, to provide a more accurate representation of the task. The subtask timings, i.e., how long it takes to open a container, were estimated via preliminary pilot tests with the experimenters as the participants. The overall workload model results significantly differed between workload conditions [22].

B. Workload Assessment Algorithm Evaluation

The workload assessment algorithm was evaluated using the physiological and subjective data collected during the evaluation. The algorithm analysis involved using six experiments. Each experiment varied the type of metric mapping and training method, i.e., neural network trained on the individual and on all participants' data (generalized). The experiments ran in real-time in order to simulate a real-world environment and generated the overall and component workload values every thirty seconds.

TABLE III

Task Weightings for the Human-Machine Teaming Experiment Note: $T\#$ represents the task number (1 - 4)

Task	**Cognitive**	**Physical**	**Auditory**	**Speech**	**Visual**
$T1_l$	0.142	0.122	0.414	0.231	0.090
$T1_h$	0.292	0.249	0.199	0.057	0.191
$T2_l$	0.074	0.354	0.358	0.191	0.023
$T2_h$	0.186	0.233	0.273	0.109	0.198
$T3_l$	0.327	0.187	0.191	0.045	0.249
$T3_h$	0.321	0.204	0.171	0.043	0.259
$T4_l$	0.177	0.558	0.103	0.025	0.135
$T4_h$	0.276	0.317	0.141	0.041	0.223

Each workload assessment algorithm experiment contained different input parameters: task weightings, physiological data, and subjective ratings. The subjective ratings consisted of only the speech and visual components. The task weightings are derived from the orignal evaluation's IMPRINT Pro model of each task [22], by dividing the average of each workload component model by the average of the overall workload model. For example, if the mean cognitive workload model value is 25 and the mean overall workload model value is 50, then the cognitive workload task weight is 0.50. The task weightings are provided in Table III, where l and h represent the low and high workload conditions, respectively.

The hypotheses are:

1) There will be a significant difference between the algorithm's workload estimates for high and low task workload conditions and high and low subjective workload conditions.
2) There will be a strong correlation between the algorithm's workload values and the workload model.
3) The normalization mapping will distinguish between workload conditions and produce more significant results than the other mappings.
4) The individually trained mappings will distinguish between workload conditions and produce higher correlations than the generalized mappings.

IV. Algorithm Analysis

There were a total of 18 low and 22 high task workload conditions analyzed. The subjective ratings determined that there was a significant difference ($\chi^2(1, N = 40) = 6.65, p = 0.01$) between low ($M = 14.05, SD = 4.14$) and high ($M = 17.65, SD = 4.00$) task workload conditions,

TABLE IV

ALGORITHMS' CALCULATED MEAN (STANDARD DEVIATION) WORKLOAD VALUES FOR ALL TASK WORKLOAD CONDITION. NOTE THE FOLLOWING ABBREVIATIONS: NORMALIZATION (NORM), REGRESSION (RNG), AND NEURAL NETWORK (NNET)

Workload Component	**Task Condition**	**Individual**			**Generalized**		
		Norm	**RNG**	**NNET**	**Norm**	**RNG**	**NNET**
Cognitive	Low	37.38 (10.85)	32.52 (18.96)	47.06 (11.28)	37.32 (10.23)	40.53 (4.83)	46.78 (11.78)
	High	36.46 (10.09)	35.31 (15.97)	47.57 (11.79)	37.99 (7.43)	40.58 (4.74)	47.92 (11.44)
Physical	Low	44.15 (7.67)	52.26 (9.27)	54.00 (4.30)	39.86 (4.81)	48.36 (1.37)	53.58 (4.69)
	High	40.96 (5.95)	48.91 (7.12)	53.28 (5.11)	40.05 (4.60)	48.08 (1.08)	53.65 (4.59)
Overall	Low	41.51 (7.21)	44.43 (6.59)	45.74 (1.89)	40.76 (6.50)	43.48 (3.52)	46.01 (4.30)
	High	43.68 (6.80)	44.37 (10.02)	46.86 (4.06)	43.40 (5.39)	45.57 (5.27)	49.05 (5.61)

TABLE V

ALGORITHMS' MEAN (STD) AND SIGNIFICANT DIFFERENCES ($\alpha = 0.05$) FOR TASK WORKLOAD CONDITIONS NOTE: $\chi^2 = \chi^2(1, N = 40)$

Training	**Mapping**	**Low**	**High**	χ^2	**P Value**
Individual	Norm	41.51 (7.21)	43.68 (6.80)	1.17	0.28
	RNG	44.43 (6.59)	44.37 (10.02)	0.18	0.67
	NNET	45.74 (1.89)	46.86 (4.06)	1.15	0.28
Generalized	Norm	40.76 (6.50)	43.40 (5.39)	4.16	**0.04**
	RNG	43.48 (3.52)	45.57 (5.27)	1.52	0.22
	NNET	46.01 (4.30)	49.05 (5.61)	3.62	0.06

TABLE VI

ALGORITHMS' MEAN (STD) AND SIGNIFICANT DIFFERENCES ($\alpha = 0.05$) FOR SUBJECTIVE WORKLOAD CONDITIONS NOTE: $\chi^2 = \chi^2(1, N = 40)$

Training	**Mapping**	**Low**	**High**	χ^2	**P Value**
Individual	Norm	40.42 (6.08)	47.61 (6.32)	11.91	**0.0003**
	RNG	42.26 (8.96)	48.82 (6.06)	5.27	**0.02**
	NNET	45.53 (3.08)	48.17 (3.23)	4.32	**0.04**
Generalized	Norm	39.95 (4.36)	47.11 (6.11)	11.91	**0.0003**
	RNG	43.04 (3.82)	48.11 (4.57)	10.55	**0.0004**
	NNET	45.93 (4.28)	51.55 (5.21)	8.93	**0.003**

TABLE VII

CORRELATION COEFFICIENTS BETWEEN ALGORITHM VALUES AND WORKLOAD MODEL NOTE: ALL VALUES ARE NON-SIGNIFICANT

Training	**Workload Component**	**Norm**	**RNG**	**NNET**
Individual	Cognitive	-0.0378	0.1225	-0.1234
	Physical	-0.0042	0.0921	-0.0146
	Overall	-0.0444	0.1548	-0.1293
Generalized	Cognitive	-0.0361	-0.0410	-0.0187
	Physical	-0.0267	-0.0367	-0.0200
	Overall	-0.0358	-0.0679	-0.0457

i.e., $T1_l - T4_l$ vs. $T1_h - T4_h$. Table IV provides the algorithms' descriptive statistics for each task workload condition and for the cognitive, physical, and overall workload components. Visual, auditory, and speech workload values are not provided, as the visual and speech workload components are static and the auditory workload component is simulated. The calculated values are separated by workload component and metric mapping. The metric mappings are categorize by training type: individual and generalized. Individual designates that the metric mapping was participant-specific, while Generalized designates that the mapping was trained on all of the participants' data.

The workload values in Table IV show that the physical workload values are higher than the cognitive workload values for each metric mapping. This result is expected, as the aggregation of the physical workload task weightings (2.224) from Table III is greater than the cognitive workload weightings (1.795). Thus, overall the tasks demanded more physical workload than cognitive workload.

The overall workload values in Table IV show that all of the metric mappings, except the participant-specific regression mapping, produced larger workload values for the high workload task condition. This result is to be expected, as the high workload tasks were designed to elicit larger overall workload values. The participant-specific regression mapping's overall workload values only differ by 0.06; thus, the mapping's overall workload values are effectively the same. It is difficult to compare task workload conditions for the cognitive and physical workload components, as the tasks were not designed to elicit larger cognitive and physical workload values for high workload conditions.

It is important that the workload assessment algorithm distinguish between low and high workload conditions. A Kruskal-Wallis test determined that there is a significant difference ($\chi^2(1, N = 40) = 4.16, p = 0.04$) between task workload conditions for the generalized normalization mapping's values. The other mappings failed to produce significant differences. The calculated p-values, group means, and standard deviations are provided in Table V.

Humans may experience different workload levels, even though a task is designed to elicit a certain workload level. Thus, a workload assessment algorithm needs to distinguish between subjective workload conditions. A Kruskal-Wallis test determined that every metric mapping resulted in a significant difference ($p \leq 0.05$) between low and high subjective workload conditions. A subjective rating above eighteen was determined to be high workload and below eighteen was determined to be low workload. Eighteen was chosen to be the threshold, as a normal workload rating (i.e., three) on each scale aggregates to eighteen. Table VI provides the statistics for the subjective workload conditions.

The algorithm's output is compared to the IMPRINT Pro's workload model in order to understand how well the algorithm tracks the expected human workload levels in real-time. The Pearson's correlation coefficient was calculated between the IMPRINT Pro model's and the algorithm's cognitive, physical, and overall workload values. The correlation coefficient for the visual, auditory, and speech workload components was not examined, since the algorithm provided static values for the workload components. Table VII provides the correlations for each training method, metric mapping, and workload component. None of the metric mappings had significant correlation coefficients. Further, seven metric mappings had a negative correlation coefficient, when the correlations were expected to be positive. The regression mapping trained on each individual's results was the only mapping that produced a positive correlation with the IMPRINT Pro workload model.

V. Discussion

The workload assessment algorithm is intended to be used by a collaborative human-machine team, where the robot adapts its interactions based on the human's workload level in real-time. The algorithm must distinguish between low and high workload conditions, in order to be viable in such a system. Hence, the first hypothesis that there will be significant differences between the algorithm's values for low and high task and subjective workload conditions was upheld. The generalized normalization mapping produced significant differences between task and subjective workload conditions. It is interesting that the normalization mapping produced significant results for the subjective workload conditions, as the mapping contained no information pertaining to the subjective ratings. However, the algorithm depended on subjective ratings for the visual and speech workload components. The normalization mapping's performance may be due to the algorithm's composition, not the metric mapping.

The second hypothesis states that there will be strong correlations between the algorithm's values and the IMPRINT Pro model. The hypothesis was not supported, as no metric mapping produced significant correlations. The poor correlations may result from an insufficient workload model. The model did not account for residual workload effects, as the model indicated multiple instances with zero workload.

The third hypothesis stated that the normalization mappings will outperform the regression and neural network mappings. The normalization mapping produced more significant values than the other two mappings. However, the generalized regression and neural network mappings produced more significant values than the generalized normalization mapping for the subjective workload conditions, which is expected, given the training used the subjective ratings. Thus, the third hypothesis is only partially supported.

The fourth hypothesis stated that the individually trained mappings will produce more significant values and higher correlations than the generalized mappings. The hypothesis was not upheld, as the generalized normalization and neural network mappings produced more significant values and higher correlations than the mappings trained on the individuals' results. The generalized regression mapping produced more significant values, but lower correlations than the individually trained regression mapping. The generalized normalization and regression mappings' high performance is attributed to the mappings having smaller group standard deviations than the individually trained mappings, which can be seen in Table V. However, the generalized neural network mapping produced higher variances than the individually trained mapping. The generalized neural network mapping's higher performance is attributed to the mapping incorporating larger mean-differences between workload conditions, than the individually trained mapping.

The workload assessment algorithm achieved the highest performance with the generalized normalization mapping, as it distinguished between task and subjective workload conditions. However, the normalization mapping had poor correlation with the IMPRINT Pro workload model. There are multiple modifications that may improve the algorithm's performance. The modifications are loosely categorized as metric weightings, variable update rates, training methods, and workload component modeling.

The current algorithm uses uniform metric weights, which negatively impact performance. For example, posture sway is an indicator of physical workload, but only if there is a large amount of torso flexing. Walking or running activities have low torso flexing, but can result in high physical workload. Thus, posture sway needs to be weighted according to the type of physical task demands. The weights need to also account for each metric's workload sensitivity. For example, heart-rate variability (HRV) is highly sensitive to cognitive workload, while skin-temperature has medium sensitivity [7]. Thus, more weight needs to be attributed to HRV.

A static thirty-second update rate is used for the workload assessment algorithm, but performance may increase with variable update rates, as sampling a workload component more frequently may provide a better workload estimate. A task with no speech component does not require a rapid speech workload component update rate. Although, a task may intermittently require speech, during which the speech component needs to update quickly to reliably capture the speech workload level. A variable update rate based on the sensor update rates and task workload components requires the use of a Kalman Filter.

The current training methods are based on static subjective ratings, but the algorithm provides continuous workload values. Training the metric mappings using continuous values can lead to higher performance. IMPRINT Pro can generate continuous workload values; however, the generated workload values do not account for individual differences.

Subjective ratings can be combined with the IMPRINT Pro model, in order to account for individual differences.

The major weakness of the current workload assessment algorithm is that only two workload components rely on physiological signals, while the other three rely on simulated signals and subjective ratings. Performance will improve if physiological metrics are used. The speech component can use speech rate, number of filler utterances and fragments, and respiration rate as physiological metrics [27], while the auditory component can use noise level and speech response time [28]. Visual workload can be assessed by using a model or physiological metrics measured via eye-tracker, i.e., blink rate and pupil dilation [29]; however, eye trackers require well-defined tracking regions.

Future work will develop a modified algorithm for collaborative human-machine teams that will allow a robot to adapt its interactions based on the human's current and projected workload levels. The presented analysis focus on peer-based teams, but the algorithm can generalize to supervisory interactions, given the algorithm's ability to assess each workload component, which permits generalization across tasks.

VI. Conclusions

Human-machine teams are capable of operating in extreme environments, where high task performance is imperative. Such teams will increase their task performance by permitting the robot to adapt its interactions based on the human's current and projected workload levels, which requires a workload assessment algorithm. Knowledge of each individual workload components' state is needed to ensure adaptations are effective, i.e., the adaptation effects the underloaded or overloaded workload component. A workload assessment algorithm is evaluated that is capable of assessing a human's overall workload state and the distinct components (cognitive, physical, auditory, speech, and visual) contributing to the overall workload state. The algorithm distinguishes between high and low workload conditions by normalizing metrics based on the entire data set.

References

[1] J. Scholtz, "Theory and evaluation of human robot interactions," in *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, 2003, pp. 10–16.

[2] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: a survey," *Foundations and Trends in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[3] C. D. Wickens, J. D. Lee, Y. Liu, and S. E. G. Becker, *An Introduction to Human Factors Engineering*, 2nd ed. Pearson Education, Inc., 2004.

[4] S. Maxwell, B. Cooper, F. Hartman, C. Leger, J. Wright, and J. Yen, "The best of both worlds: Integrating textual and visual command interfaces for mars rover operations," in *IEEE International Conference on Systems, Man and Cybernetics*, 2005, pp. 1384–1388.

[5] J. McCraken and T. Aldrich, "Implications of operator workload and system automation goals," U.S. Army Research Institution, Tech. Rep. ASI-479-024-84B, 1984.

[6] P. Jorna, "Heart rate and workload variations in actual and simulated flight," *Ergonomics*, vol. 36, no. 9, pp. 1043–1054, 1993.

[7] M. Castor, *GARTEUR Handbook of Mental Workload Measurement*, ser. GARTEUR technical publications. Group for Aeronautical Research and Technology in Europe, 2003.

[8] H. A. Abbass, J. Tang, R. Amin, M. Ellejmi, and S. Kirby, "Augmented cognition using real-time EEG-based adaptive strategies for air traffic control," in *Human Factors and Ergonomics Society Annual Meeting*, vol. 58. SAGE Publications, 2014, pp. 230–234.

[9] K. T. Durkee, S. M. Pappada, A. E. Ortiz, J. J. Feeney, and S. M. Galster, "System decision framework for augmenting human performance using real-time workload classifiers," in *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision*, 2015, pp. 8–13.

[10] J. Aasman, G. Mulder, and L. Mulder, "Operator effort and the measurement of heart-rate variability," *Human Factors*, vol. 29, no. 2, pp. 161 – 170, 1987.

[11] K. Vicente, D. C. Thornton, and N. Moray, "Spectral analysis of sinus arrhythmia: A measure of mental effort," *Human Factors*, vol. 29, no. 2, pp. 171–182, 1987.

[12] T. C. Hankins and G. F. Wilson, "A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight," *Aviation Space and Environmental Medicine*, vol. 69, no. 4, pp. 360–367, 1998.

[13] A. Roscoe, "Assessing pilot workload. Why measure heart rate, HRV and respiration?" *Biological Psychology*, vol. 34, no. 2-3, pp. 259–287, 1992.

[14] J. Keller, H. Bless, F. Blomann, and D. Kleinbohl, "Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol," *Journal of Experimental Social Psychology*, vol. 47, no. 4, pp. 849–852, 2011.

[15] S. Miyake, S. Yamada, T. Shoji, Y. Takae, N. Kuge, and T. Yamamura, "Physiological responses to workload change. a test/retest examination," *Applied Ergonomics*, vol. 40, no. 6, pp. 987–996, Nov 2009.

[16] T. Mizuno, T. Sakai, S. Kawazura, H. Asano, K. Akehi, S. Matsuno, K. Mito, Y. Kume, and N. Itakura, "Measuring facial skin temperature changes caused by mental work-load with infrared thermography," *IEEE Transactions on Electronics, Information and Systems*, vol. 136, no. 11, pp. 1581–1585, 2016.

[17] M. Brenner, E. T. Doherty, and T. Shipp, "Speech measures indicating workload demand," *Aviation, Space, and Environmental Medicine*, vol. 65, no. 1, pp. 21–26, 1994.

[18] D. Harris, *Human Performance on the Flight Deck*. Ashgate Publishing Limited Surrey, U.K., 2011.

[19] P. Paul, F. M. Kuijer, B. Visser, and H. C. G. Kemper, "Job rotation as a factor in reducing physical workload at a refuse collecting department," *Ergonomics*, vol. 42, no. 9, pp. 1167–1178, 1999.

[20] D. Lasley, R. Hamer, R. Dister, and T. Cohn, "Postural stability and stereo-ambiguity in man-designed visual environments," *IEEE Transactions on Biomedical Engineering*, vol. 38, no. 8, pp. 808–813, 1991.

[21] J. A. Johnstone, P. A. Ford, G. Hughes, T. Watson, and A. T. Garrett, "Bioharness multivariable monitoring device: Part. ii: Reliability," *Journal of Sports Science & Medicine*, vol. 11, no. 3, p. 409, 2012.

[22] C. E. Harriott, "Workload and task performance in human-robot peer-based teams," Ph.D. dissertation, Vanderbilt University, 2015.

[23] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," *Journal of the American College of Cardiology*, vol. 37, no. 1, pp. 153–156, 2001.

[24] C. M. Bishop, "Model-based machine learning," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, pp. 2012– 2222, 2012.

[25] S. Archer, M. Gosakan, P. Shorter, and J. Lockett, "New capabilities of the Armys maintenance manpower modeling tool," *Journal of the International Test and Evaluation Association*, vol. 26, no. 1, pp. 19 – 26, 2005.

[26] A. International, "Standard guide for operational guidelines for initial response to a suspected biothreat agent," American Society for Testing and Materials, techreport ASTM E2770-10, 2010.

[27] A. Berthold and A. Jameson, "Interpreting symptoms of cognitive load in speech input," in *Conference on User Modeling*, J. Kay, Ed. Springer Vienna, 1999, pp. 235–244.

[28] J. B. Clark and C. S. Allen, "Acoustics issues," in *Principles of Clinical Medicine for Space Flight*. Springer Nature, 2008, pp. 521–533.

[29] B. Cain, "A review of mental workoad literature," Defence Research and Development Toronto, techreport RTO-TR-HFM-121-Part-II, 2007.