Measuring Variations in Workload during Human-Robot Collaboration through Automated After-Action Reviews

Zhiqin Qian[*]
Rice University
Houston, TX, USA
bill.qian@rice.edu

Liubove Orlov Savko[*]
Rice University
Houston, TX, USA
liuba.orlov.savko@rice.edu

Catherine Neubauer
Army Research Laboratory
Aberdeen Proving Ground, MD, USA
catherine.e.neubauer2.civ@army.mil

Gregory Gremillion
Army Research Laboratory
Aberdeen Proving Ground, MD, USA
gregory.m.gremillion.civ@army.mil

Vaibhav Unhelkar
Rice University
Houston, TX, USA
vaibhav.unhelkar@rice.edu

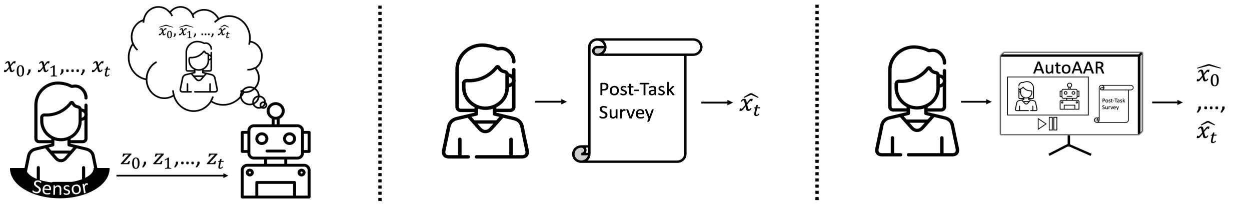

Figure 1: Measuring temporal variations in workload during human-robot collaboration is difficult using existing methods. (left) Objective measures provide objective but indirect measures $(z_0, \dots, z_t)$. (middle) Subjective measures derived using post-task surveys typically only provide aggregate measures (x_t). (right) This paper presents AutoAAR: a method to extract direct time series measurements of workload by enhancing existing post-task surveys using concepts from after-action reviews.

ABSTRACT

Human collaborator's workload plays a central role in human-robot collaboration. Algorithms designed to minimize cognitive workload enhance fluent human-robot teamwork. Time series data of workload is vital for both designing and assessing these algorithms. However, accurately quantifying and measuring cognitive workload, particularly at high temporal resolution, poses a substantial challenge. Towards addressing this challenge, we explore the potential of after-action reviews (AARs) as a tool for gauging workload during human-robot collaboration. First, through a case study, we present and demonstrate AutoAAR for measuring human workload post-task at a high temporal resolution. Second, through a user study, we quantify the validity and utility of measurements derived using AutoAAR for human-robot teamwork. The paper concludes with guidelines and future directions to extend this method to measure other internal states, such as trust and intent.

CCS CONCEPTS

• **Human-centered computing**; • **Computer systems organization** → **Robotics**; • **Computing methodologies** → Artificial intelligence;

KEYWORDS

Methods, Data Collection, Human Internal States

ACM Reference Format:
Zhiqin Qian, Liubove Orlov Savko, Catherine Neubauer, Gregory Gremillion, and Vaibhav Unhelkar. 2024. Measuring Variations in Workload during Human-Robot Collaboration through Automated After-Action Reviews. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3610978.3640677

1 INTRODUCTION

The success of fluent collaboration between humans and robots often hinges on a variety of human internal states, including intent, workload, and trust in the robot. Several algorithmic approaches have been developed to estimate human intent, optimize cognitive workload, foster trust in robots, and align robot behavior with human values [3, 8, 24, 33, 35, 36]. Studies have indicated that when these states are overlooked, robots exhibit suboptimal behaviors, resulting in less effective human-robot teamwork [5, 32].

Design of these techniques relies on data of human behavior, including time series of their internal states. However, as depicted in Fig. 1, measuring these temporal variations is difficult using existing methods. Acknowledging the inherent difficulty in measuring human internal states, a branch of research focuses on partially observable decision-making techniques for human-robot collaboration [8, 17, 19, 21, 28, 29, 34]. These techniques do not measure internal states directly; instead, they utilize sensor observations to estimate and react to them. However, even these partially observable techniques need data of internal states to assess their efficacy.

[*] Both authors contributed equally to this research.

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3640677

Recognizing this need, this paper focuses on methods to measure human's cognitive workload (a key internal state) during human-robot collaboration (HRC). As workload is an intrinsically latent quantity, measuring it is a complex endeavor. Methods need to fulfill competing objectives of being

(R1) accurate (i.e., generate accurate measurements);
(R2) temporally sensitive (i.e., capture temporal variations);
(R3) non-intrusive (i.e., not impact task execution);
(R4) user-friendly (i.e., easy to implement and use).

Related Methods. Principled methods developed to measure human workload include physiological measures [4, 14, 18, 26] and self-report questionnaires [13, 31]. As depicted in Fig. 1, each method presents unique advantages and challenges. Physiological methods are objective but do not offer direct measurements of workload and additional processing is required to estimate workload from physiological data [1, 9, 12, 38]. Post-hoc methods like NASA-TLX [13] are non-intrusive but do not provide measurements at a high temporal resolution. Further, these methods provide non-contemporaneous measurements, which can introduce additional memory and perception bias. In contrast, self-report probes administered during task execution can be disruptive to task performance and thus the cognitive states that they aim to measure [10].

Summary of Contributions. To complement existing methods, this paper aims to contribute to the methodological toolkit available to HRI researchers. We present AutoAAR: a self-report method for gathering measurements of workload. Not only is AutoAAR non-intrusive, but it also provides measurements at a high temporal resolution, a key requirement for algorithmic HRI. Sections 2–3 introduce AutoAAR and demonstrate its use in a simulated HRC task. Sections 4–5 discuss a user study, compares AutoAAR with other methods, and confirms its utility for human-robot collaboration.

2 METHOD

Automated After-Action Review (AutoAAR) is a self-report method that seeks to fulfil the competing requirements (introduced in Sec. 1) for measuring temporal variations in workload during human-robot collaboration. It builds on existing self-report questionnaires and enhances them to be temporally sensitive and non-intrusive by incorporating features of after-action reviews.

Background. AutoAAR is inspired by the practice of after-action review, which is a structured debriefing method used for performance improvement [2, 27]. It involves a human trainer together with the trainee(s) examining the events that occurred during a task, understanding the reasons behind them, and determining how to improve performance in the future. These structured reviews have found a few applications in human-machine interaction [6, 16, 25, 30], most focusing on making behavior of artificial agents more transparent. Instead, we use after-action reviews to understand behavior of their human counterparts in human-machine teams. Unlike traditional after-action reviews, AutoAAR is tailored for subjective workload measurement and incorporates automated components, enabling its use without the need for a human trainer.

Prerequisites. For effective use of AutoAAR, a video replay of the task is crucial. We foresee several potential applications of AutoAAR in near-term HRI scenarios, especially those in semi-structured settings like factories, offices, or classrooms as well as in human-robot training settings, such as research labs or simulation environments. However, AutoAAR may not be suitable for contexts where recording and replaying the interaction is either untenable (e.g. due to privacy concerns) or impractical.

Measurement Workflow in AutoAAR. In AutoAAR, the human teammate is prompted to review the execution of a human-robot collaborative task immediately after its completion. This review process is supported by a video replay of the task. AutoAAR, through an interactive user interface, asks the participant to answer self-report questions on their workload. By administering these questions post-task any interference with the task execution is prevented (cf. R3). Further, by administering these questions periodically during the review, AutoAAR is capable of capturing temporal variations in workload (cf. R2). To counteract potential errors stemming from memory and perception biases, AutoAAR integrates mechanisms inspired by after-action reviews to aid recall and minimize reporting fatigue, ensuring more accurate and reliable reporting.

Steps for Administering AutoAAR. To effectively utilize AutoAAR, we recommend the following procedure:

(1) determine self-report scales for measuring workload;
(2) set up mechanisms to record videos of the HRC task;
(3) create a user interface that replays the task execution and periodically administers the self-report scales;
(4) include mechanisms to facilitate recall; and
(5) incorporate mechanisms to reduce reporting fatigue, such as optimizing the number of questions through pilot trials.

3 CASE STUDY

We now present a case study that illustrates AutoAAR in action for the simulated rescue domain: Rescue World for Teams (RW4T) [20]. The RW4T simulation considers a dyadic team composed of a human first responder and a semi-autonomous robot drone. The team is tasked with delivering multiple first-aid kits in a radioactive area under time constraints, necessitating human-robot collaboration. While both agents are capable of dispensing first aid kits, the human needs to designate goal locations for the robot. The robot moves autonomously given a goal but, informed by practical considerations, exhibits imperfect behavior and fails to distribute a kit at the intended location in 30% of instances. The human also needs to respond to secondary tasks of varying intensity during task completion. For more details of this task and environment, please refer to [20]. Here, we describe the process of designing the AutoAAR interface using the five steps outlined in Sec. 2 for this HRC task.

Step 1: Self-Report Questionnaire. We selected the following prompt, "Indicate your level of workload," administered on a 5-point Likert scale to measure the participant's cognitive workload.

Step 2: Mechanisms to Record Task Execution. To construct a task replay, we captured the state of the environment in every simulation frame during task execution. Given that the task was simulated, recording both the state of the task and the actions of the team was feasible; for real-world tasks, AutoAAR would necessitate the use of recording equipment such as cameras or motion capture systems.

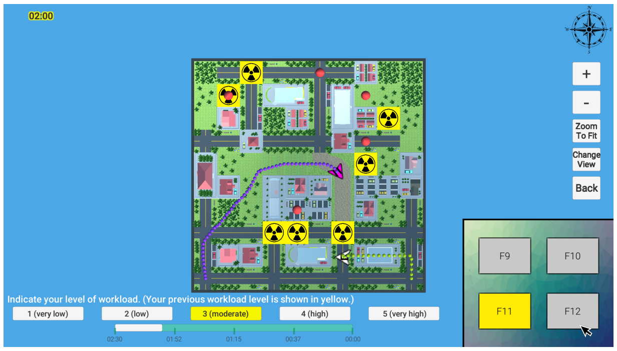
Indicate your level of workload. (Your previous workload level is shown in yellow.)

Figure 2: Snapshot of the AutoAAR interface.

Step 3: Mechanisms for Playback and Review. To facilitate the replay of recorded task execution, we developed an interactive user interface using the Unity game engine shown in Fig. 2.[1] The logged data enabled a precise replay to be reconstructed. Upon task completion, the interface displayed the replay to the human teammate and paused it at periodic 30-second intervals to inquire about their workload at those specific moments during the collaborative task.

Step 4: Mechanisms to Facilitate Recall. Similar to after-action reviews, AutoAAR must include mechanisms to facilitate recall [30]. Our implementation included three such mechanisms. First, the user interface displayed trails of the human and robot during the replay. Second, during AutoAAR, the user interface allowed the human teammate to view the domain from multiple perspectives. Third, the user interface included the following recall questions for each reporting period: 1) "How many drop-off locations did you visit?", and 2) "How many drop-off locations did the robot visit?". These questions were also administered at 30-second intervals along with the question for measuring workload.

Step 5: Mechanisms to Reduce Reporting Fatigue. Like other self-report methods [27, 37], AutoAAR must include mechanisms to reduce reporting fatigue. In our implementation, the frequency of probing was selected based on pilot trials to strike a balance between obtaining data with high temporal granularity (cf. R2) and avoiding making the data labeling process overly tedious (cf. R4). Further, when probing their workload, the interface also showed their previous answers as a point of reference. If the participant reported a value that differed from their previous response, they were additionally asked the following question: "You indicated a change in workload since the last period. Please use the arrow keys to rewind to the moment when this change occurred." Using such change-points, AutoAAR increases the granularity of the measurements without increasing the frequency of self-report probes.

4 USER STUDY

We conducted a user study to assess the validity and utility of measurements derived using AutoAAR.[2] In particular, through this study, we explore the following research questions:

(Q1) How do workload measurements from AutoAAR compare with those derived via established self-report methods?
(Q2) How do workload measurements derived from AutoAAR compare with those derived via physiological sensors?
(Q3) Can workload measurements from AutoAAR be used to improve human-robot collaboration?

The user study was conducted using the RW4T simulator task described in Sec. 3, where participants served the role of human teammate. The study utilized two physiological sensors: Zephyr BioHarness and Tobii Pro Nano Eye Tracker. The experiment protocol was approved by Rice University's IRB. After giving informed consent, participants completed a demographic survey, which included queries about their age, gender, and video game experience. Participants were then instructed to wear and follow the calibration procedures of the physiological sensors, which included a 3-minute waiting period to establish baseline physiological measurements and enable cross-subject comparisons. To acquaint participants with the simulation and experimental interface, an initial tutorial session and a training trial were provided.

Subsequently, participants engaged in four test trials of the HRC task outlined in Sec. 3. The environmental layout was consistent across trials; however, the drop-off and starting locations varied. To introduce variability in workload, the secondary task's intensity was altered across the four test trials. Each trial was segmented into four periods, presenting either low or high intensity of secondary task. Low-intensity periods had no secondary tasks, while high-intensity periods required participants to engage in a secondary task every 3 seconds. The participants were unaware in advance of when secondary tasks would occur. Each participant received a compensation of $12, with a bonus of $12 awarded to the participant achieving the best teaming performance with the robot.

We estimate participant's cognitive workload via three methods: AutoAAR, NASA-TLX (an established self-report method), and physiological sensors.[3] First, during the task, we collect time series data $(z_1, z_2, \ldots)$ of the following physiological measures using the BioHarness and Eye Tracker: *heart rate, heart rate variability, blink frequency, blink duration, blink latency, pupil dilation, and fixation rate.* Second, we administered an abridged NASA-TLX survey after the completion of each task trial. This provides an aggregate measure of workload (i.e., $\hat{x_t}$) for each trial. Third, upon completing the NASA-TLX, participants were asked to perform AutoAAR for that trial. This review utilized the interface developed in Sec. 3 to derived time series data $(\hat{x}_1, \hat{x}_2, \ldots)$ of workload.

5 EXPERIMENTAL RESULTS

We now report results of the user study, which included 24 participants (4 female, median age: 21 years).

Q1. Consistency Between AutoAAR and Subjective Measures. NASA-TLX is an established scale for measuring workload; however, it provides aggregate measures but not high-frequency time-series data. Hence, to evaluate the internal consistency, we compute correlations between NASA-TLX measurements and the average of time series data derived using AutoAAR. We find that the AutoAAR measurements have a Pearson correlation of 0.46 with the

[1] A video demonstration of the interface is available at http://tiny.cc/autoaar-video
[2] To facilitate reproducibility, we provide additional details regarding the user study via an Appendix available at http://tiny.cc/autoaar-appendix
[3] Lab Streaming Layer was used to synchronize data collected from different sources. To facilitate analysis, the data streams were resampled at 1Hz frequency.

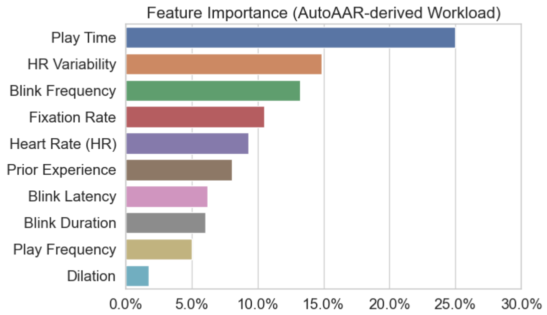

Figure 3: Feature importance of physiological and game-playing features in predicting AutoAAR-derived workload.

unweighted average NASA-TLX and a correlation of 0.41 with its weighted counterpart. These correlations suggest *moderate degree of agreement* between these two measurement methods.

Q2. Consistency Between AutoAAR and Objective Measures. Physiological sensors do not provide direct workload measurements, making direct comparisons with AutoAAR infeasible. Hence, to evaluate their internal consistency, we assess the *predictability* of AutoAAR-derived workload using physiological data. Heard et. at [14] outlined various predictive models for estimating workload from physiological data. Drawing upon these studies, we train models to predict AutoAAR-derived workload from physiological measures, using predictability as a surrogate for internal consistency. Among the models we tested,[4] random forests yielded the most accurate results and was able to predict AutoAAR-derived workload with over 97% accuracy, *suggesting high-degree of internal validity.* We also find that incorporating personalization (via game-playing features) improved accuracy. Figure 3 illustrates the normalized contribution of each feature in predicting workload. Both physiological and game-playing features play a role in the prediction, with the demographic features together accounting for 38% feature importance. This exploration of feature importance underscores the crucial role of recognizing individual differences in cognitive state assessments and further hints at AutoAAR 's adeptness in reconciling these variances.

Q3. Utility for Human-Robot Teamwork. We next assess the utility of AutoAAR-derived measurements for enhancing human-robot collaboration. Previously, methodologies ranging from rule-based function allocation (like MABA-MABA) to planning algorithms have been developed to generate robotic behavior that depends on human workload [7, 11, 23]. Informed by these works and to answer Q3, we simulate workload-dependent robotic assistance: if a participant's workload exceeds a set threshold, the robot assists with secondary tasks, affecting the team score. This proof-of-concept analysis is conducted retrospectively, using human subject data collected in the user study.

Table 1 reports the increase in team performance (denoted as score increment), number of robot interventions (#), and performance gain per intervention (denoted as efficiency). We observe that using AutoAAR-derived workload measures (both actual and predicted values) facilitates efficient and timely robot interventions, wherein the robot intervenes only when necessary. Recall that the predicted value of AutoAAR-derived workload can be calculated during task execution, by leveraging a trained model and sensor data. Given this, the observed efficiency of AutoAAR (predicted) condition is especially encouraging, as it demonstrates that AutoAAR-derived data can be used to generate adaptive robot behavior that enhances human-robot collaboration.

Workload	Score	Interventions	
measure:	Increment	Total #	Efficiency
AutoAAR (reported)	31.7	6.7	4.76
AutoAAR (predicted)	31.4	6.6	**4.79**

Table 1: Effect of workload-based robotic interventions, computed using different workload measures, on team score.

6 CONCLUSION AND FUTURE DIRECTIONS

This paper introduces AutoAAR, a method for collecting time-series data of workload from individuals engaged in human-robot tasks, without disrupting their activities or necessitating wearable sensors. Using a visual task replay and leveraging validated questionnaires, AutoAAR is able to satisfy the competing requirements for workload measurements introduced in Sec. 1. We confirm the validity and utility of AutoAAR-derived measurements of workload through a user study ($N = 24$).

Due to its various features, we believe that it can serve as a method of choice for measuring workload variations in domains where recording and reviewing the HRC task is feasible. Researchers and practitioners seeking to use AutoAAR should follow the 5-step procedure outlined in Sec. 2 and illustrated in the case study of Sec. 3. Our investigation and its limitations also motivate several directions of future work. First, our experiments have been confined to a disaster response scenario. While we observed positive results, further validation through replication studies is imperative. Second, our experiments occurred in a simulated environment, motivating further work to assess its applicability in non-simulated settings.

Third, our work merely begins to explore the extensive possibilities of after-action reviews within human-robot collaboration. Beyond workload, AutoAAR can be adapted to measure other time-series data reflecting human internal states like intent, engagement, and trust in robots. Its non-intrusive, cost-effective attributes make AutoAAR an ideal tool for simulation test beds, facilitating the collection of human behavior data with annotations of human internal states. While further investigation and methodological advancements are necessary to fully realize these applications, our findings lay the groundwork for the broader adoption of after-action review principles in future human-robot interaction research.

ACKNOWLEDGMENTS

This research was supported by the Army Research Office through CA# W911NF-20-2-0214. Liubove Orlov Savko was supported in part by the Fulbright-García Robles fellowship.

[4] For more details on model training and selection, please refer to the Appendix available at http://tiny.cc/autoaar-appendix

REFERENCES

[1] Muneeb Imtiaz Ahmad, Ingo Keller, David A Robb, and Katrin S Lohan. 2020. A framework to estimate cognitive load using physiological data. *Personal and Ubiquitous Computing* (2020), 1–15.

[2] Gary Allen and Roger Smith. 1994. After action review in military training simulations. In *Proceedings of Winter Simulation Conference*. IEEE, 845–849.

[3] Mara Baljan and Peter Nickel. 2019. *Level of Robot Autonomy and Information Aids in Human-Robot Interaction Affect Human Mental Workload – An Investigation in Virtual Reality*. 278–291. https://doi.org/10.1007/978-3-030-22216-1_21

[4] Cindy L Bethel, Kristen Salomon, Robin R Murphy, and Jennifer L Burke. 2007. Survey of psychophysiology measurements applied to human-robot interaction. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 732–737.

[5] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 708–713.

[6] Ralph Brewer, Anthony Walker, E. Pursel, Eduardo Cerame, Anthony Baker, and Kristin Schaefer. 2019. *Assessment of Manned-Unmanned Team Performance: Comprehensive After-Action Review Technology Development*. 119–130. https://doi.org/10.1007/978-3-030-20467-9_11

[7] Claudia Carissoli, Luca Negri, Marta Bassi, Fabio Alexander Storm, and Antonella Delle Fave. 2023. Mental Workload and Human-Robot Interaction in Collaborative Tasks: A Scoping Review. *International Journal of Human–Computer Interaction* (2023), 1–20.

[8] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2020. Trust-Aware Decision Making for Human-Robot Collaboration: Model Learning and Planning. *ACM Transactions on Human-Robot Interaction* 9 (02 2020), 1–23. https://doi.org/10.1145/3359616

[9] Essam Debie, Raul Fernandez Rojas, Justin Fidock, Michael Barlow, Kathryn Kasmarik, Sreenatha Anavatti, Matt Garratt, and Hussein A Abbass. 2019. Multimodal fusion for objective assessment of cognitive workload: a review. *IEEE transactions on cybernetics* 51, 3 (2019), 1542–1555.

[10] Francis T Durso, Carla A Hackworth, Todd R Truitt, Jerry Crutchfield, Danko Nikolic, and Carol A Manning. 1998. Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly* 6, 1 (1998), 1–20.

[11] Paul M Fitts. 1951. Human engineering for an effective air-navigation and traffic-control system. (1951).

[12] Yao Guo, Daniel Freer, Fani Deligianni, and Guang-Zhong Yang. 2021. Eye-tracking for performance evaluation and workload estimation in space telerobotic training. *IEEE Transactions on Human-Machine Systems* 52, 1 (2021), 1–11.

[13] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[14] Jamison Heard, Caroline E Harriott, and Julie A Adams. 2018. A survey of workload assessment algorithms. *IEEE Transactions on Human-Machine Systems* 48, 5 (2018), 434–451.

[15] Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. 2008. Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 symposium on Eye tracking research & applications*. 69–72.

[16] H. Lane, Mark Core, Michael Lent, Steve Solomon, and Dave Gomboc. 2005. Explainable Artificial Intelligence for Training and Tutoring. In *Artificial Intelligence in Education*. 762–764.

[17] Mikko Lauri, David Hsu, and Joni Pajarinen. 2022. Partially observable markov decision processes in robotics: A survey. *IEEE Transactions on Robotics* 39, 1 (2022), 21–40.

[18] Catherine Neubauer, Kristin E Schaefer, Ashley H Oiknine, Steven Thurman, Benjamin Files, Stephen Gordon, J Cortney Bradford, Derek Spangler, and Gregory Gremillion. 2020. *Multimodal Physiological and Behavioral Measures to Estimate Human States and Decisions for Improved Human Autonomy Teaming*. Technical Report. CCDC Army Research Laboratory Aberdeen Proving Ground United States.

[19] Liubove Orlov-Savko, Abhinav Jain, Gregory M Gremillion, Catherine E Neubauer, Jonroy D Canady, and Vaibhav Unhelkar. 2022. Factorial Agent Markov Model: Modeling Other Agents' Behavior in presence of Dynamic Latent Decision Factors. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS.

[20] Liubove Orlov-Savko, Zhiqin Qian, Gregory M Gremillion, Catherine E Neubauer, Jonroy Canady, and Vaibhav Unhelkar. 2024. RW4T Dataset: Data of Human-Robot Behavior and Cognitive States in Simulated Disaster Response Tasks. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[21] Stefania Pellegrinelli, Henny Admoni, Shervin Javdani, and Siddhartha Srinivasa. 2016. Human-robot shared workspace collaboration via hindsight optimization. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 831–838.

[22] Mary C Potter, Brad Wyble, Carl Erick Hagmann, and Emily S McCourt. 2014. Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics* 76 (2014), 270–279.

[23] Matthew S Prewett, Ryan C Johnson, Kristin N Saboe, Linda R Elliott, and Michael D Coovert. 2010. Managing workload in human–robot interaction: A review of empirical studies. *Computers in Human Behavior* 26, 5 (2010), 840–856.

[24] Peizhu Qian and Vaibhav V Unhelkar. 2022. Evaluating the Role of Interactivity on Improving Transparency in Autonomous Agents. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS.

[25] Jennifer Riley and Mica Endsley. 2004. The Hunt for Situation Awareness: Human-Robot Interaction in Search and Rescue. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 48 (09 2004). https://doi.org/10.1177/154193120404800389

[26] Raphaëlle Roy, Nicolas Drougard, Gateau Thibault, Frédéric Dehais, and Caroline Ponzoni Carvalho Chanel. 2020. How Can Physiological Computing Benefit Human-Robot Interaction? *Robotics* 9 (11 2020), 100. https://doi.org/10.3390/robotics9040100

[27] Margaret S Salter and Gerald E Klein. 2007. *After action reviews: current observations and recommendations*. Technical Report. WEXFORD GROUP INTERNATIONAL INC VIENNA VA.

[28] Sangwon Seo, Bing Han, and Vaibhav Unhelkar. 2023. Automated Task-Time Interventions to Improve Teamwork using Imitation Learning. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS.

[29] Sangwon Seo and Vaibhav Unhelkar. 2024. IDIL: Imitation Learning of Intent-Driven Expert Behavior. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS.

[30] Michael Taberski, Kristi Davis, Kristin Schaefer, and Ralph Brewer. 2021. *Visualizing Human-Autonomy Team Dynamics Through the Development of a Global After-Action Review Technology*. 46–53. https://doi.org/10.1007/978-3-030-79763-8_6

[31] Andrew J Tattersall and Penelope S Foord. 1996. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics* 39, 5 (1996), 740–748.

[32] Vaibhav V Unhelkar, Przemyslaw A Lasota, Quirin Tyroller, Rares-Darius Buhai, Laurie Marceau, Barbara Deml, and Julie A Shah. 2018. Human-Aware Robotic Assistant for Collaborative Assembly: Integrating Human Motion Prediction with Planning in Time. *IEEE Robotics and Automation Letters (RA-L)* 3, 3 (2018), 2394–2401.

[33] Vaibhav V Unhelkar, Shen Li, and Julie A Shah. 2020. Decision-Making for Bidirectional Communication in Sequential Human-Robot Collaborative Tasks. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[34] Vaibhav V Unhelkar, Shen Li, and Julie A Shah. 2020. Semi-supervised learning of decision-making models for human-robot collaboration. In *Conference on Robot Learning*. PMLR, 192–203.

[35] Sebastijan Veselic, Claudio Zito, and Dario Farina. 2020. Human-Robot Interaction with Robust Prediction of Movement Intention Surpasses Manual Control. *bioRxiv* (2020). https://doi.org/10.1101/2020.12.09.416735 arXiv:https://www.biorxiv.org/content/early/2020/12/11/2020.12.09.416735.full.pdf

[36] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Vienna, Austria). ACM, 408–416. https://doi.org/10.1145/2909824.3020230

[37] Qiping Zhang, Austin Narcomey, Kate Candon, and Marynel Vázquez. 2023. Self-Annotation Methods for Aligning Implicit and Explicit Human Feedback in Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 398–407.

[38] Yueying Zhou, Shuo Huang, Ziming Xu, Pengpai Wang, Xia Wu, and Daoqiang Zhang. 2021. Cognitive workload recognition using EEG signals and machine learning: A review. *IEEE Transactions on Cognitive and Developmental Systems* (2021).

APPENDIX

A AUTOAAR INTERFACE IMPLEMENTATION

Extending the simulated rescue domain introduced in [20], we implemented the AutoAAR interface using the Unity game engine. While the human-robot team was engaged in the task, we captured data in every frame of the simulation to construct a replay. The data captured included the positions of the human, the robot, and the remaining first aid kits, as well as the team's actions, like task-relevant keystrokes and button presses. The logged data enabled precise reconstruction of each simulation frame, allowing an exact replay of the collaborative task.

During the replay, the following questions were administered at each pause during AutoAAR:

- How many drop-off locations did you / robot visit during this period?
- Indicate your level of workload.
- (If change in reported workload) You indicated a change in workload since the last period. Please use the arrow keys to rewind to the moment when this change occurred.
- Open-ended comments (optional).

B USER STUDY

B.1 Participant Recruitment

We recruited 24 participants through campus flyers and departmental mailing lists. Their ages ranged from 18 to 46, with a median of 21. Among of the participants, 5 identified as female.

B.2 Demographic Survey

Questions administered during demographic survey:

- Age
- Gender
- What is your prior experience with video games?
- How often do you play video games?
- On average, how much time do you spend each time you play a video game (in minutes)?
- Open-ended comments (optional).

B.3 Physiological Measurements

B.3.1 BioHarness. Using BioHarness, we collect data of:

- *Heart rate.* The rate of heart beats per minute. It is calculated at a frequency of 1Hz.
- *Heart rate variability.* The degree of variation between heart beats. It is also calculated at a frequency of 1Hz.

B.3.2 Tobii Eye Tracker. The Tobii eye tracker keeps track of the position of the left and right eyes on a computer screen as well as the diameter of each pupil. If the eye gaze cannot be captured, i.e. when participants blink or look off the screen, the eye tracker will output NaN instead. To detect blinks as accurately as possible, we did not count NaN values as blinks if they happened concurrently with keyboard presses (F9, F10, F11, F12) for secondary tasks, as participants tend to look down at the keyboard to find the secondary task keys. Using the eye tracker, we collect data of:

- *Blink frequency.* The number of blinks in a time window (30 seconds in our case unless otherwise specified).
- *Blink duration.* The average duration of the blinks (in seconds) in a time window.
- *Blink latency.* The average duration between blinks (in seconds) in a time window.
- *Pupil dilation.* The average diameter of the pupils. This is collected by the eye tracker at a rate of 60Hz. We performed smoothing on the raw data using a low pass filter, following the recommendation by Klingner et al. [15].
- *Fixation rate.* The number of fixation changes in a time window, where fixation is when one gazes at an Area of Interest (AOI) for a preset period of time around 30ms). This is chosen as Potter et. al found that humans can detect meaning in images in as little as 13 milliseconds [22].

	Degree	Learning Rate
AutoAAR workload (w/ game)	3	0.0001
AutoAAR workload (w/o game)	3	0.1

Table 2: Best hyperparameters for each type of polynomial regression model. "w/ game" means that the model uses game-playing features, and "w/o game" means otherwise.

	Maximum Depth
AutoAAR workload (w/ game)	23
AutoAAR workload (w/o game)	24

Table 3: Best hyperparameters for each type of random forest model. "w/ game" means that the model uses game-playing features, and "w/o game" means otherwise.

Batch Size	Learning Rate	Dropout Rate	Nodes per Layer
21	0.001	0.1	256

Table 4: Best hyperparameters for neural network models.

C CONSISTENCY BETWEEN AUTOAAR AND OBJECTIVE MEASURES

To evaluate internal consistency of AutoAAR-derived and physiological measures, we assess the predictability of AutoAAR-derived workload using physiological data. Supervised learning was deployed to develop predictive models $f(z, \theta) = x$, where z represents the various physiological measures, x represents workload, and θ represents video game-playing features derived from the demographic survey. An 80/20 training/test split was applied.

Predictive models were trained using three techniques: polynomial regression, random forest, and neural networks. Hyperparameters were selected using grid search and cross-validation. The hyper-parameter search space is listed as follows:

- Polynomial Regression
  - *degree*=$\{1, 2, 3, 4, 5\}$
  - *learning rate*=$\{0.1, 0.01, 0.001, 0.0001\}$
- Random Forest
  - *number of decision trees*=$\{100\}$
  - *maximum depth*=$\{15, 16, 17, \cdots, 25\}$
- Neural Network
  - *batch size*=$\{32, 64\}$
  - *learning rate*=$\{0.01, 0.001\}$
  - *dropout rate*=$\{0.1, 0.2\}$
  - *nodes per hidden layer*=$\{128, 256\}$

The best values of hyper-parameters (amoung our search space) for each of the three supervised learning techniques are listed in Tables 2–4.