Online Human Workload Detection for Behavior Adaptation in Autonomous Robot Teammates*

Ayca Aygun[1][0000−0001−8157−2219], Helena Fu[1], Evan Krause[1], and Matthias Scheutz[1][0000−0002−0064−2789]

[1]Tufts University, Medford MA 02155, USA
ayca.aygun@tufts.edu

Abstract. Managing human cognitive load, especially in teams, is important for ensuring high task performance. We hypothesize that robots which can monitor human cognitive load in real-time in mixed initiative teams and adapt their interactions based on it will lead to better team performance compared to robots that are unaware of human cognitive load. In this paper, we introduce an online cognitive workload detection algorithm based on changes in human pupil size that is able to track human cognitive workload during task performance. The algorithm is integrated into a cognitive robotic architecture to allow the system to adapt its behavior based on inferred workload. The system is evaluated in a mixed initiative human-robot team experiment where two humans and two autonomous robots had to collaborate in order to achieve high performance in a complex dual-task setting. The results confirm the operation of the proposed integrated system and demonstrate that online adaptation of autonomous robot behavior to human cognitive load can lead to better team performance compared to non-adaptive robots.

Keywords: Robot Autonomy, Real-Time Cognitive State Estimation, Cognitive Workload, Dynamic Systems, Adaptive Systems, Pupillometry, Human-Robot Teaming

1 Introduction

As robots are gaining more autonomy, they are increasingly envisioned as teammates in mixed-initiative teams where humans and robots need to collaborate to accomplish common goals, with application domains ranging from search and rescue scenarios on Earth, to space missions like NASA's planned cislunar Gateway station as part of NASA's Moon-to-Mars strategy. Especially in high-demand environments like rescue or space missions, managing human cognitive workload to sustain task performance while ensuring the safety and well-being of human participants is critical. Adaptive autonomous systems that can respond to human mental states in real-time, in particular, human cognitive workload, thus have the potential to enhance overall team effectiveness. Even though there have been research efforts that explore how robots can respond to human workload, it is still unclear whether real-time cognitive workload detection and real-time adaptation of robot behavior in response to it will be able to improve team performance by reducing human stress.

In this paper, we introduce a novel algorithm for online human workload detection based on changes in human pupil size which has previously been shown to be an effective and accurate way of offline estimation of human cognitive workload (*e.g.*, [2]). Pupillometry has several advantages

* This work was funded in part by AFOSR grants #FA9550-18-1-0465 and #FA9550-23-1-0425.

over other sensory modalities that have been used for detection cognitive workload such as EEG or fNIRS in that it does not have any motion artifacts, is easy to collect, and allows humans to operate in their environment freely (instead of being seated or tethered to measurement equipment, for example). The real-time workload detection algorithm is then integrated into a cognitive robotic architecture to provide the robot with moment-by-moment estimates of human cognitive workload which allows the robot to adapt its behavior based on workload levels. To validate the integration and to demonstrate that adapting to cognitive workload can be effective and improve team performance, we evaluate the integrated system in a complex mixed-initiative team settings where two human and two autonomous robot teammates have to accomplish a multi-objective team task. The results from the empirical study show that human workload tracking and subsequent adaptation of robot behavior can lead to significant performance improvements, verifying to our knowledge for the first time experimentally the utility of adjusting robot behavior dynamically based on human cognitive needs. The outcome points to an important capability for future robots deployed in human-machine teams to improve team effectiveness.

2 Related Work

Several recent research efforts have focused on understanding human cognitive workload in a dynamic autonomous human-robot interaction settings. [15] proposed a cognitive workload prediction method during robot-assisted surgery using physiological sensing and machine learning methodologies. While it showed promising results for estimating cognitive workload, it solely focused on *passive workload monitoring* while our proposed framework actively integrates real-time workload assessment into the robot architecture, enabling dynamic behavioral adaptation of robots during collaborative tasks. Other work aimed at understanding the utility of eye tracking measures for estimating mental workload across different user groups and task conditions primarily focused on offline analysis [13] while we propose a real-time method for assess cognitive workload and dynamically adapting robot behavior during task execution. Moreover, while [13] emphasized individual performance, our experimental setup involves collaborative team-based interactions with multiple human and robot agents, offering more complex and valid contexts for evaluating adaptive HRI. Similarly, [12] provided a thorough investigation into sensitivity and reliability of eye tracking metrics for assessing cognitive workload during physical human-robot co-manipulation tasks, but the primary focus was on offline mental workload analysis across different task difficulty levels. Additionally, their study is centered on individual motor learning while our experimental paradigm involves multi-agent collaboration (two human and two artificial agents) in a complex, dynamic environment, emphasizing team performance and adaptive interaction over time. [7] presented an approach to workload management in a multimodal HRI platform using deep reinforcement learning framework to incorporate both subjective (self-reported) and objective (physiological) cognitive workload measures for dynamic task reallocation. In contrast, our approach uses only real-time eye gaze data for workload adaptation, offering a simple, continuous, and fully non-intrusive solution for estimating cognitive load. Some efforts have also investigated combined physiological measures such as [14] who developed a multi-sensing, domain-specific adaptive automation system for robotic-assisted surgery, relying on both EEG and eye tracking data. In contrast, our pupil-only approach eliminates the need for intrusive, complex sensing setups involving EEG, making it more practical for diverse, real-world HRI contexts beyond surgery. And some efforts for real-time workload adaptation rely on broader agent condition monitoring and explicit workload transitions such as [8] who proposed a modular workload allocation framework for multi-human multi-robot systems,

emphasizing task reallocation based on both human and robot health/performance in real-time. In contrast, our work focuses specifically on real-time cognitive workload estimation using only eye gaze, providing a simple workable solution for adaptive support in multimodal HRI scenarios and can be used both for task-level as well as within-task adaptations.

3 Online Pupillometry-Based Cognitive Workload Detection

We start by introducing our proposed algorithms for cognitive workload detection using changes in pupil size and then discuss in the next section how it is used to adapt robot behavior. Since pupil size is in general affected by ambient light, it is important to distinguish changes due to light alone from changes due to the influence of cognitive processes. Moreover, since people exhibit individual variations in pupil size and pupil size changes, it is important to determine for each individual their range of pupil size changes as well as their response to light hitting the retina. We thus start with a description of how to account for these two critical factors by devising a careful calibration routine which allows us to quickly build an individual's pupil model that can later be used during task performance to measure cognitive workload.

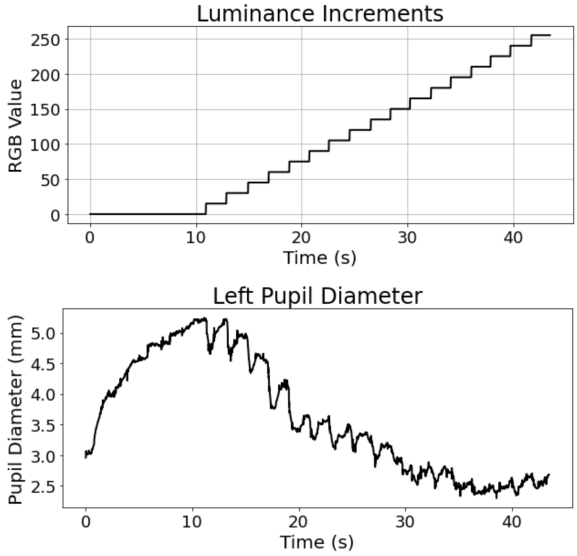
3.1 Building Pupil Size Change Models

To measure workload using pupil size, an eye-tracker is needed that can detect and track the changes in pupil diameter which, however, can be influenced by varying lighting conditions that can impact the accuracy of the measurements [9]. In fact, pupil size is primarily influenced by changes in luminance whereas the effect of cognitive workload on pupil diameter is comparatively small [10]. As a result, it is crucial to address luminance variations, either through experimental design or data processing, to prevent them from obscuring the impact of cognitive workload. We thus developed a luminance elimination technique that allows us to develop a model of pupil size changes for each individual using the following calibration routine. Note that we have used a VR headset for this routine as it allowed us fully control the light hitting the participant's retina by way of generating images that are displayed in the headset; but it is also possible to do this without the headset if ambient light can be fully controlled in the participant's location.

In our calibration routine using VR headsets, participants are first shown a black screen (*i.e.*, complete darkness), followed by a gradual transition of the whole screen towards a completely white screen in 18 incremental steps. The grayscale values range from 0 (black) to 255 (white) across these increments. The initial black screen duration lasted for 10 seconds, allowing the participants' pupils to fully adjust to the darkness, as pupil response time requires some acclimatization. Each subsequent increment lasts for 2 seconds, gradually increasing the grayscale value. The luminance calibration ends when the screen reaches full white (RGB value of 255). Figure 1 shows the incremented values in time and the impact of the luminance screen incrementation on the left pupil size.

After all pupil size values are obtained, we fit a polynomial model to the data. In particular, the function takes an array including the pupil size responses associated with the incremented luminance screen values (as seen in Figure 1) and generated a polynomial function that models the relationship between the luminance effect and the pupil size in time. We used the "polyfit" function **np.polyfit** from the NumPy Python library to generate the best fitting cubic polynomial for the pupil size data. This function generates the coefficients for the polynomial of order 3. Figure 2 shows the fitted model of the participants' pupils' response to the various screen luminance values which can

Fig. 1. Luminance increments (top) and the participant's left pupil response to the luminance effect (below).



Algorithm 1 Pupil data pre-processing

Require: Raw pupil gaze data G

function PREPROCESSPUPIL(G) ▷ **Pre-processing of raw eye signal**
  for $i = 1$ to $|G|$ **do**
    if $G[i] < 0.8$ or $G[i] > 10$ **then**
      $G[i] \leftarrow$ NaN
    end if
  end for
  return Interpolated(G) ▷ **Linear interpolation**
end function

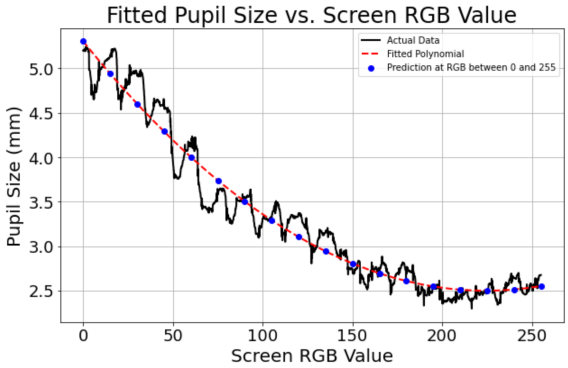
be then used to determine what changes in pupil size one can expect for the particular individual given any amount of ambient light. Any differences between the expected luminance-based pupil size changes and the actual changes are then due to additional factors like cognitive effort and can thus be used for estimating cognitive load.

3.2 Determining Sense of Urgency Threshold

In order to be able to track and predict workload levels, we have to ensure that the eye gaze signals are valid and do not have unrealistic pupil size values (which can sporadically happen due to sensor errors). Hence, we perform an initial simple denoising procedure (see Algorithm 1) to remove physiologically implausible pupil size values, *i.e.*, signals lower than 0.8 mm which are potential blink artifacts and signals greater than 10 mm which are unrealistic [1] and perform linear interpolation to fill in the missing data points.

Next, we need to determine individual pupil-based thresholds for each human team member to detect moments of high cognitive effort for that individual during task performance, *i.e.*, instances of a *sense of urgency* [3] which, when exhibited frequently over a given time period, have been

Fig. 2. The **np.polyfit** function generated by using the pupil size values associated with the incremented screen luminance values.



shown to increase cognitive workload (which develops over a longer period of time than fast pupil size changes). This is an essential preparatory step that can be performed in a short "proxy task" to collect individual pupil size and luminance values as the effects of task demands on an individual's cognitive system and thus the manifestation in pupil size changes can be different. Any task can be used as a proxy task here as long as it has varying cognitive demands, ideally starting with no cognitive demand and ramping up the effort to "unachievable levels" in order to determine an individual's performance ceiling.

The overall sense-of-urgency threshold estimation Algorithm 2 then takes a time series of raw pupil sizes G, the corresponding luminance levels of the environment L when the pupil size was recorded and the previously determined luminance-based pupil size estimation function f_{lum} and removes the estimated luminance impact on pupil size from the preprocessed pupil signals to generate luminance-corrected pupil signals that is smoothed using a moving average filter. The smoothed signal is then squared to compute its power, amplifying larger fluctuations indicative of cognitive effort. A second moving average is applied to the power signal to further reduce noise and emphasize sustained changes. Finally, the sense of urgency threshold τ is calculated as the mean of this smoothed power signal plus two standard deviations, a common statistical technique for identifying significant outliers. The rationale for using the power of the signal to detect peaks is that significant changes or features in the signal are often accompanied by substantial increases in power which captures the signal's energy and strength, making it an effective measure for identifying cognitive effort which might be due to a sense of urgency [2]. The threshold then is based on identifying values that are significantly higher than the average which, in a normal distribution, fall beyond two standard deviations from the mean are considered rare (approximately 5% of the data). The equation related to threshold calculation can be seen as follows:

$$\tau = \mu_{\tilde{P}} + 2 \times \sigma_{\tilde{P}} \tag{1}$$

where $\mu_{\tilde{P}}$ is the mean and $\sigma_{\tilde{P}}$ is the standard deviation of the smoothed power signal.

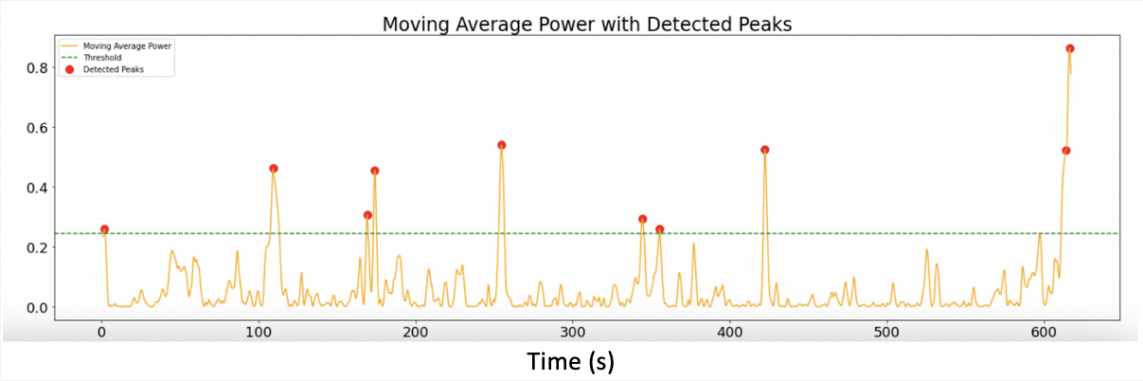
Note that the threshold provides an objective marker of effort-related peaks in the pupil data (see Figure 3) rather than just relying on arbitrary cutoff values and can thus be used for real-time workload monitoring by integrating the frequency of sense of urgency events for that particular individual over a period of time in subsequent tasks.

Algorithm 2 Workload Threshold Estimation (G, L, f_{lum})

Require: Raw pupil gaze data G, Luminance values L, and Luminance pupil size function f_{lum}

$G^* \leftarrow \text{PREPROCESSPUPIL}(G)$ ▷ **Filter raw pupil signal**
for $i = 1$ to $|L|$ **do**
  $\hat{G}[i] \leftarrow G^*[i] - f_{\text{lum}}(L[i])$ ▷ **Luminance correction**
end for
$\tilde{G} \leftarrow \text{movingaverage}(\hat{G})$ ▷ **Moving average of** $\hat{G}$
$P \leftarrow \tilde{G}^2$ ▷ **Power of the eye gaze**
$\tilde{P} \leftarrow \text{movingaverage}(P)$ ▷ **Moving average of** P
$\tau \leftarrow \mu_{\tilde{P}} + 2\sigma_{\tilde{P}}$ ▷ **Threshold calculation**
return τ

Fig. 3. An example of moving average power of the pupil size signal with detected peaks based on the threshold.



3.3 Real-time Workload Prediction

Real-time workload prediction can then be accomplished for an individual after determining the individual's luminance pupil size function f_{lum} and sense-of-urgency threshold τ by repeatedly running Algorithm 3. The algorithm performs the same initial steps as the sense-of-urgency threshold estimation Algorithm 2, but instead of returning τ, it identifies the maximum value within the smoothed power signal of the pupil size and returns whether it exceeds τ. Note that it is important for the algorithm to work correctly to select a sufficient long window of observation; if it is too short, the algorithm will not be able to detect workload *per se* but only sense-of-urgency moments. We have found experimentally that a window of 30 seconds is indicative for heightened cognitive effort. To evaluate the utility of the above methods, Algorithm 3 needs to be integrated into a robotic architecture in a way that allows the robot to use the workload (high/low) signal to adapt its behavior.

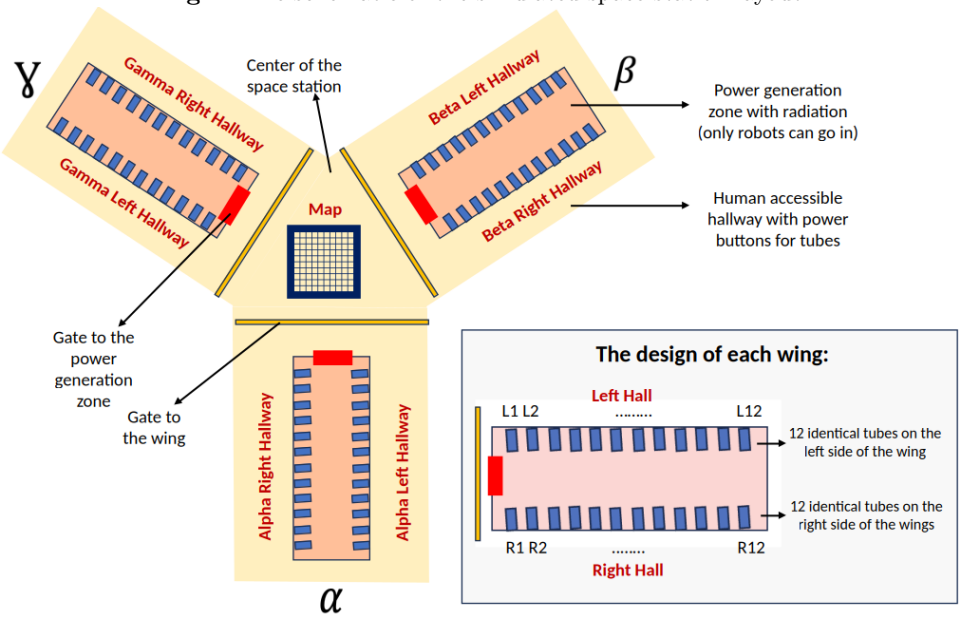
4 Experimental Methods

For the experimental evaluation of the utility of our online workload detection method, we decided on a within-subject design where we could directly compare a workload-based adaptive with a non-adaptive condition. We utilized an existing Unity3D simulation environment that provided

Algorithm 3 Real-time workload prediction (G, L, f_{lum}, τ)

Require: Raw pupil gaze data G, Luminance values L, Luminance pupilsize function f_{lum}, and threshold τ
$G^* \leftarrow \text{PREPROCESSPUPIL}(G)$
for $i = 1$ to $|L|$ **do**
    $\hat{G}[i] \leftarrow G^*[i] - f_{\text{lum}}(L[i])$ ▷ **Adjust for luminance**
end for
$\tilde{G} \leftarrow \text{movingaverage}(\hat{G})$ ▷ **Moving average of eye gaze**
$P \leftarrow \tilde{G}^2$ ▷ **Calculation of power of eye gaze**
$\tilde{P} \leftarrow \text{movingaverage}(P)$ ▷ **Moving average of power signal**
$a \leftarrow \max(\tilde{P})$ ▷ **Take the maximum value of $\tilde{P}$**
return $a > \tau$ ▷ **Return true for high workload, false otherwise**

Fig. 4. The schematic of the simulated space station layout.



the infrastructure for running two humans and two fully autonomous robots in a complex mixed-initiative team task which consisted of two tasks, one non-collaborative task which only humans could perform, and one collaborative task which could only be performed by humans and robots working together. The latter was important for measuring the utility of workload-based adaptive robot behavior, while the former was used for putting cognitive load on the human participants. We will first describe the experimental setup with the details of the two tasks, followed by the differences of the robots' behaviors in the two conditions and the details of the experimental procedure.

4.1 The Space Station Mixed-Initiative Multi-Human Multi-Robot Team Task

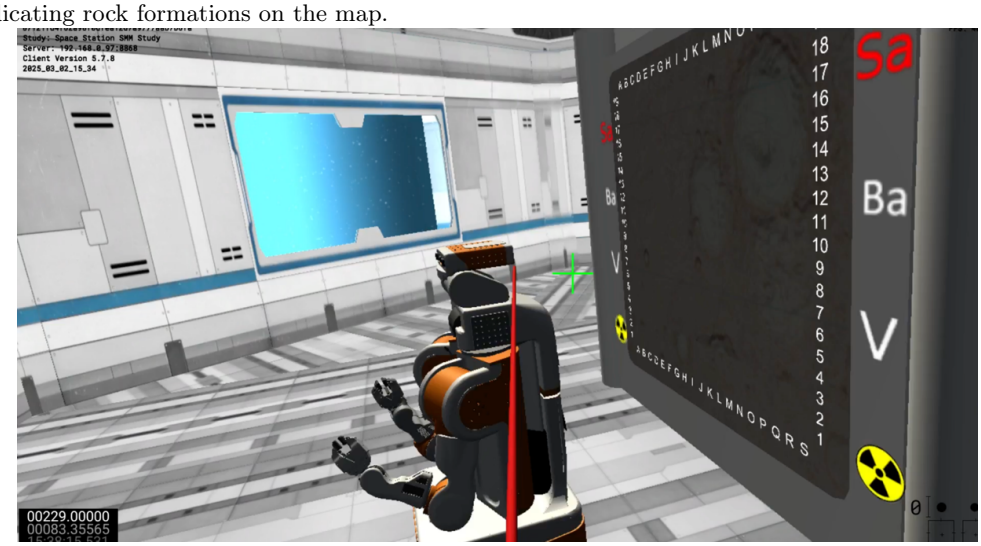
The Unity3D environment is that of a simulated future NASA-inspired Space Station orbiting Mars[1] which has been successfully used in previous user studies for evaluating human-robot interactions

[1] The complete code for the simulation environment is available here https://github.com/mscheutz/diarc/wiki/DIARC-Unity-Space-Station-Simulation.

in multi-task collaborative team settings (*e.g.*, for evaluating the utility of shared mental models [5, 6]).

Figure 4 depicts the layout of the simulated space station with three identical wings labeled as "Alpha", "Beta", and "Gamma" and a central area with a map. Each wing has a yellow outer gate through which human participants and robots can enter the wing, a red gate located inside the wing that leads to a power generator area which only robots can enter due to the heightened radiation, and two hallways (left and right) through which human teammates can view the power generation area from outside. There are 24 identical power tubes inside each power generation area (12 on the left and 12 on the right side) which provide power to the station and contribute to the station's overall health. Every so often tubes are getting damaged due to continued use and need to be repaired. The **joint human-robot repair task** consists of three steps: (1) a human participant must press the power on/off button in the hallway outside the damaged tube to turn it off; (2) a robot must move to the damaged tube inside the power generation area to repair it; and (3) the tube needs to be turned on again by a human participant to be able to generate power for the station. Damaged tubes continue to deteriorate and produce less and less energy depending on the damage, and if they are not repaired in time, they will permanently break at which point they can no longer be fixed. The goal for the **joint human-robot repair task** is to keep the station in the best possible operating condition, *i.e.*, at the maximum health level possible.

Fig. 5. One side of the map and one robot in the central area of the space station as see though the eyes of a human teammate (the green cross is the participant's eye gaze, the red rod is the participant's effector for indicating rock formations on the map.



The center of the space station hosts a three-sided map with rows numbered 1 to 19 vertically and columns labeled A to S horizontally where humans need to indicate rock formations (*i.e.*, **Sandstone** (Sa), **Basalt** (Ba), **Volcanic** (V), and **Radiation**) detected by a rover on the surface of Mars. The rover transmits coordinates and types of formation verbally at different times and the task for a human – the **rover map task** – is to indicate them in the proper location on the map

(repair robots are not needed or involved this task). The goal of the **rover map task** is to indicate all communicated formations correctly on the map. Figure 5 shows one side of the map and one robot from the view of a human participant.

4.2 System Setting and Robot Behavior

The experiment platform running the Unity3D space simulation included a simulation server for running Unity as well as a Windows computer that works with Steam VR to connect to an HTC Vive Pro Virtual Reality headset, equipped with eye tracking capabilities. Simulated PR2 robots in the Space Simulation were connected to DIARC robotic architecture instances running on separate computers. The robots were controlled through spoken voice commands. To interact with a robot, participants needed to press and hold one of the two triggers on their controllers while speaking. Their audio was then transmitted to Kaldi (an open-source speech recognition toolkit) on an automatic speech recognition (ASR) server to transcribe the speech. The transcription was sent to the DIARC architecture. In turn, robots communicate with human teammates using utterances generated by the DIARC natural language generator (NLG) component and synthesized by OpenTTS which converts them into audio files that are played as the robots' voices in the participants' headsets.

Since the Unity3D space station setting already provided the DIARC architecture [11] with a standard robot monitoring behavior, we utilized the behavior for the **non-adaptive condition**. In this condition, two robots move through from wing to wing in the space station to check each wing for broken tubes. If a broken tube is found, the robots notify the human participants that a tube is broken in the respective wing. However, the robots do not specify which tube is broken, only informing them that there is one or more broken tubes in a particular wing while the move on to the next wing. Robots will interrupt their monitoring behavior when instructed by humans to go to a particular wing to repair a particular tube ("go to Alpha and repair tube left four"). After the repair, they automatically resume their monitoring behavior.

For the **adaptive condition**, we first integrated the workload detection algorithms into the DIARC architecture (which was used to run both robots) and the developed a modified action script based on the above behavior as follows: Every 30 seconds the workload of both human participants was estimated, and as long as they were both below the individual τ threshold value for each participant, the robot continued to perform the above monitoring behavior. Once the workload estimate went above the individual threshold for one human teammate, the robot switched to adaptive mode and remained in adaptive mode for thirty seconds. In adaptive mode, once it detects a damaged tube, it announces the wing *and* the number of the damage tube to prompt a human teammate to go there and turn it off, so the tube can be repaired before it permanently breaks (different from the non-adaptive condition where the robot only announces the wing). The robot then waits by the damaged tube for a predetermined time, reminding the humans in between that it is waiting to do the repair (at the expense of failing to notice tubes breaking in other wings). If no human shows up to turn the tube off, the robot continues with the monitoring routine when new workload measures are below the threshold, otherwise it continues to remind the humans about the damaged tube (again, at the expense of reporting any damaged tubes in other wings). Hence, while the adaptive robots explicitly inform humans of damaged tubes and position themselves by those tubes to make it easier for humans to remember to turn them off and on again, they are not able to perform their monitoring task at that time, and thus will not be able to detect and report newly damaged tubes, which is a critical *tradeoff* between aiding the repair based on workload adaptation compared to continuing to spot failing tubes between the adaptive and non-adaptive condition.

4.3 Procedure

After participants arrived, they were seated in separate rooms and were asked to sign the consent form. Once the consent form was signed, participants received instructions about the experiment and the procedure, followed by setting up the VR headset and performing the calibration process as described in Section 3.1. After completing the calibration, the participants were instructed to explore the spaceship and were given guidance on how to use the controllers for both navigation and interaction with the objects. The next phase involved a 2-minute practice trial for practicing the rover map task, during which participants focused solely on indicating rock formations communicated by the remote rover on the map. Then another 2-minute practice trial was added to allow participants to practice repairing broken tubes where they had to interact with one of the robots. During this trial, we also provided guidance on how to communicate with the robots, giving participants sample transcriptions. Some examples of the possible command are:

- Go to alpha
- Come to beta
- Check for broken tubes in gamma
- What tubes in beta are broken
- Which tubes in alpha are broken
- Go to alpha right five
- Repair gamma left one

The final preparatory part was a 5-minute trial of a single-person single-robot version of the experiment where participants were asked to complete both rover map and tube repair tasks. This trial was used as the "proxy task" to determine the sense-of-urgency threshold τ (as described in Section 3.2). Specifically, we examined participants' workload individually by gradually increasing the frequency of rover communications and of tubes getting damaged as the trial progressed (14 rover communications and 11 damaged tubes). This increment in event frequency was intended to induce an increasingly higher workload on participant and allowed us to determine the individual τ thresholds for use in the subsequently two-person two-robot runs.

After completing the preparatory phase, both participant were allowed to remove their headset and relax for a couple of minutes. Once they had settled, we informed them that the next phase would involve working with two robotic agents and one other human teammate to accomplish the same tasks. We also informed them that they were free to coordinate their activities in any way they chose with their human teammates. However, we refrained from providing any additional guidance on how they should coordinate, so as not to influence their natural teaming behaviors.

The experiment consisted of two sequential 10-minute trials: one adaptive and one non-adaptive. The order of these trials was counterbalanced to avoid ordering effects as we expected the possibility of training effects (as the initial training was comparatively short for a complex task setting like this). Participants could communicated with each other through a headset (as they were seated in separate rooms) which allowed us to slightly distort their voices to avoid any incidental identification of the teammate by voice (this was done to ensure that there would not be any significant difference in teaming behavior due to the acquaintance of the participants).

At the beginning of each 10-minute trial, we set up the VR environment and performed eye tracking and luminance calibration to ensure accurate eye gaze tracking. Each 10-minute, 2-human 2-robot trial included 24 tube repairing tasks and 45 rover mapping tasks. To induce a higher workload, we gradually increased the frequency of event occurrences throughout each trial (as with the individual calibration trials).

4.4 Participants

We recruited 48 participants from Tufts University and conducted 24 trials, each involving two participants. We had to eliminate the data from eight trials due to issues with the robots and VR connection during the experiment. We also had to exclude three additional pairs because the workload levels of either participant never surpassed their threshold τ in the adaptive condition which for them was the same as the non-adaptive condition. Hence, we had 13 usable trials (recorded data from 26 participants in total). 59% of the participants identified themselves as male, while the remaining 41% identified themselves as female. A majority (84%) were right-handed. The whole study lasted at most 1.5 hours. The study was approved by our university's Institutional Review Board (IRB). Each participant provided informed consent and completed a guided tutorial before the start of the experiment. Upon completion of the whole experiment, participants were compensated with $20 for their participation.

5 Experimental Results

To investigate the differences between adaptive and non-adaptive two-person, two-robot scenarios, we used two sets of three objective measurements each, one set pertaining to the joint human-robot task of repairing damaged tubes, and another pertaining just to the rover task performed only by the human teammates. For the first, we measured the number of *broken tubes* and *repaired tubes* as well as *overall station health* (as the sum of the energy contributed by all tubes based on their level of operation) at the end of each trial; for the second, we measured *correct* and *incorrect* indications of rock formations on the map as well as *missed* ones (see Table 1 for the average values of these measurements across all 13 experiments).

Table 1. Comparison of the average values of the six objective measures in the adaptive and non-adaptive trials, with the top three measuring human-robot team performance.

Measure	**Non-Adaptive**	**Adaptive**
Mean # broken tubes	12.23	9.23
Mean # repaired tubes	4.46	8.54
Mean final station health	37.08	47.92
Mean # correct markings	16.62	20.23
Mean # incorrect marking	1.54	1.46
Mean # missed placements	28.00	24.08

Since the data was not balanced, we could not perform mixed-effects ANOVA analyses, but rather used R's lme4 library [4] to perform linear mixed effects analyses for all six measures. Here we report in detail the three human-robot team performance measures that are directly affected by the robot autonomy condition and only summarize the outcomes of the other three. Specifically, we conducted linear mixed-effects analysis to examine the effects of adaptive vs. non-adaptive autonomy and order of the workload conditions on repaired tubs, with random intercepts for participants. The model showed that the non-adaptive condition resulted in a significant decrease in the number of repaired tubes compared to the adaptive condition, ($\beta = -4.737$, $SE = 1.240$, $t(11) = -3.821$, $p = 0.00284$). The order in which conditions were presented also had a significant effect on the outcome,

with participants showing overall higher tube repair in the second run ($\beta = 2.863$, $SE = 1.240$, $t(11) = 2.309$, $p = 0.04137$) which suggests a training effect. The random effects analysis revealed significant variability in the intercepts across participants (variance = 0.9377, standard deviation = 0.9683), indicating that individual differences in baseline outcomes were substantial. The residual variance was 9.4580 (standard deviation = 3.0754) suggesting that a considerable portion of the variability in outcomes remained unexplained by the model. The degrees of freedom for each of the fixed effects were approximated using Satterthwaite's method, and the reported values reflect this adjustment. The model fit was assessed using restricted maximum likelihood (REML), with a REML criterion at convergence of 126.1.

We also conducted a linear mixed-effects model analysis to examine the effect of condition (adaptive vs. non-adaptive) on the number of broken tubes, with random intercepts for participants. The model revealed participants in the non-adaptive condition had a significantly higher number of broken tubes than those in the adaptive condition, ($\beta = 3.000$, $SE = 0.906$, $t(11) = 3.312$, $p = 0.0062$). There was no order effect. The model included random intercepts for participants, which showed significant variability in baseline outcome scores across participants (variance = 0.1923, standard deviation = 0.4385). The residual variance was 5.3333 (standard deviation = 2.3094), suggesting that there is considerable unexplained variability in the outcome. The model fit was assessed using restricted maximum likelihood (REML), with a REML criterion of 114.2 at convergence.

Finally, we also conducted a linear mixed-effects model to examine station health at the end of each trial with the same independent variables of autonomy condition and order and obtained a significant effect of condition, with participants in the non-adaptive condition reporting significantly lower station health scores compared to those in the adaptive condition ($\beta = -11.700$, $SE = 3.114$, $t(11) = -3.757$, $p = 0.00317$). Order was again not significant. Random effects analysis showed variability in the baseline health outcomes across participants (variance = 0.2795, standard deviation = 0.5287). The residual variance was 59.6909 (standard deviation = 7.7260), indicating considerable unexplained variability in health outcomes. The model fit was assessed using restricted maximum likelihood (REML), with a REML criterion of 166.4 at convergence.

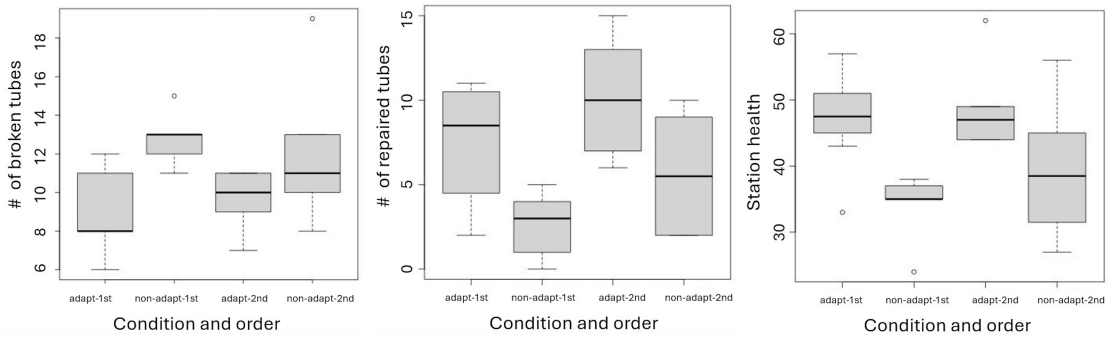
Overall, the results confirm that adaptive condition was helpful for the participants, given that the task was complex and stressful, and led to a significantly higher team performance compared to non-adaptive robots (see also Figure 6). Robots spent on average 80% of their time in adaptive mode in the adaptive condition. But note that this does not necessarily translate to a difference in robot behavior compared to a non-adaptive robot unless the robots spots a broken tube.

For the human-only rover task (where robots could only indirect help by monitoring wings for broken tubes and thus allowing humans to focus on the map) we found only one significant order effect indicating a reduction in missed rover communications and thus placements of markings in the second run compared to the first, suggesting again a training effect, but we did not find any effects on correctly or incorrectly, or even missed markings (even though numerically the results point in the direction of improvements in the adaptive condition).

6 Discussion, Limitations, and Future Work

We performed the first experimental evaluation of real-time pupil-based workload detection and dynamic robot behavior adaptation in a mixed-initiative team that goes beyond human-robot dyads with two humans and two language-enabled fully autonomous robots. The results we presented here already point to the utility of an important capability that is currently missing in most autonomous

Fig. 6. Comparison of adaptive vs non-adaptive trial performance across the three metrics related to the joint human-robot repair task.



robots: The ability to track human cognitive load and adapt the robot's behavior to provide support when needed. In our experiments, this support came in the form of the robot's interruption of its monitoring and reporting behavior in exchange for a proactive movement to the broken tube and a verbal reminder to the human teammates that it was ready to repair this particular tube as soon as somebody turned it off. The downside of this interruption was that robots missed newly damaged tubes. But it had the advantage of reminding people who were at that time overloaded (e.g., because they were attending to the map task, or they were attempting to navigate through the environment as quickly as possible) to perform their part of the joint task (*i.e.*, to turn off the damaged tube) which they otherwise might have forgotten given everything that was going on. In the non-adaptive condition some participants never found out which tubes were damaged or they forgot to attend to them, even though they were regularly informed about wings with newly broken tubes. Hence, the verbal reminder in the adaptive condition as well as the robot's readiness allowed human participants to be more efficient in repairing damaged tube and preventing them from breaking completely. The significant differences in all three objective measures in the joint human-robot tube repair task between the adaptive and non-adaptive conditions of *the same team*, therefore, is evidence that the tradeoff in robot behaviors to actively support human actions as opposed to continuing with their individual monitoring task and possibly detecting newly damaged tubes paid off–at a certain point it is more important to repair a tube than to know about what other tubes are broken. And the adaptive behavior we selected for the robots implicitly underscored this point. But note that there was nothing in the adaptive condition that could not have happened in the non-adaptive condition with respect to robot actions: Participants could, and did, summon robots to damaged tubes. And also note that the robot adaptation in the adaptive condition was not *per se* a better behavior as performing it for the whole time would have resulted in a worse overall outcome and thus lowered team performance. I.e, had the robots in the adaptive condition completely dropped their monitoring task (*e.g.*, in exchange for sticking with a damaged tube that needed to be repaired instead of resuming monitoring when human workload was low), the human teammates would not even have learned about damaged tubes in other wings, let alone been able to go there in time to turn them off for repair. Overall, the experiment shows that properly chosen adaptive robot tasks that support human teammates when their performance is likely impacted by high workload can lead to better performance even when this means that the robot will have to interrupt another important task. But the nature of the adaptation will essentially depend on the

particulars of the task setting and cannot be generalized across tasks and environments, or teaming arrangements.

This also raises an important limitation with this type of study regarding the nature of the adaptation. Some adaptations are clearly *maladaptive* in that they lower performance, and we could have also added such a condition where the robot, for example, based on detecting high human workload would have simply stopped in place without doing anything else unless prompted. Such a condition would have resulted in even worse performance than the non-adaptive condition because such robots would not even have noticed broken tubes during conditions of high human workload, leaving that task to their human teammates as well. A human with a high workload then would likely remain in that state due to the additional task for monitoring for broken tubes, causing the robot to remain stationary unless explicitly instructed to perform repairs, a vicious circle. We did not include this condition because the goal of the study was to show that a *helpful adaptation* triggered by high workload was able to reduce workload and lead to better team performance. But while the added information of the particular broken tube as well as the robot's proactive movement to the tube was helpful, it is not always clear at the outset of a task which behavior might have positive effects. Hence, it is likely that robots in teams will need to have additional mechanisms for tracking the effects of potential adapations they are capable of and selecting the ones that are most conducive for improving team performance in response to high human workload.

As the main focus of this paper was to introduce the adaptive workload detection algorithm and evaluate its utility for online robot behavior adaptation in mixed-initiative teams, there are several aspects of the experiment that cannot be included in the analyses and discussion here and have to await a future publication. Specifically, it would be interesting to analyze the communications between the two human teammates to determine whether and to what extent they discussed and agreed on a division of labor. For example, we have seen participant pairs where both participants did both the map and the repair tasks, while for other participant pairs, one participant is entirely in charge of the map task while the other is entirely in charge of the repair. This division of labor, for example, reduces coordination effort in that only one person needs to interact with the robots (since no robot interaction is needed for the map task). Yet, it comes at the expense of failing to repair some of the tubes before the break because it is not always physically possible to move quickly enough between wings where tubes are failing (in those cases the former teaming arrangement is better suited). In addition to the coordination of the human participants, we see some teams working with fixed robot assignments to human partner compared to others where every participant can command any of the two robots. All of these variations and their effects on team effectiveness would be important to analyze, especially with respect to the interactions with and impact of adaptive robot of team performance. These analyses will also require additional experimental runs in order to get enough (unprompted) examples of the different arrangements and coordination strategy.

7 Conclusion

The aim of this paper was twofold: (1) to introduce an individual-based novel method for real-time cognitive workload detection using pupil size changes indicative of sense-of-urgency events over a given window of time, and (2) to evaluate the algorithm integrated into a cognitive robotic architecture in a complex human-machine teaming setting with two humans and two autonomous robots who must perform two types of tasks, one that can only be performed by humans, the other that can only be performed by humans and robots working together. Experimental evaluations of 13 participant pairs in an immersive simulated space station environment in Unity3D showed that

robots which are able to adapt their behavior based on human workload estimates can significantly improve objective task performance measures for the joint human-robot task. This suggests for future mixed-initiative human-machine teams that if human eye gaze data is already available or can be easily collected (which is typically possible without negatively impacting human actions), it can serve as a useful additional input channel for robots to determine when to change their behaviors to actively support human performance and lower their workload which then, ultimately, can benefit the whole team.

References

1. Aygun, A., Lyu, B., Nguyen, T., Haga, Z., Aeron, S., Scheutz, M.: Cognitive workload assessment via eye gaze and eeg in an interactive multi-modal driving task. In: Proceedings of the 2022 international conference on multimodal interaction. pp. 337–348 (2022)
2. Aygun, A., Nguyen, T., Haga, Z., Aeron, S., Scheutz, M.: Investigating methods for cognitive workload estimation for assistive robots. Sensors **22**(18), 6834 (2022)
3. Aygun, A., Nguyen, T., Scheutz, M.: Assessment of multiple systemic human cognitive states using pupillometry. In: Proceedings of the 46th Annual Meeting of the Cognitive Science Society (2024)
4. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. Journal of Statistical Software **67**(1), 1–48 (2015). https://doi.org/10.18637/jss.v067.i01
5. Edgar, G., Aygun, A., McWilliams, M., Scheutz, M.: Towards genuine robot teammates: Improving human-robot team performance beyond shared mental models with proactivity. In: Vinjamuri, R.K. (ed.) Emerging Frontiers in Human-Robot Interaction. Springer Nature (2024)
6. Gervits, F., Thurston, D., Thielstrom, R., Fong, T., Pham, Q., Scheutz, M.: Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In: Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) (2020)
7. Jo, W., Wang, R., Yang, B., Foti, D., Rastgaar, M., Min, B.C.: Cognitive load-based affective workload allocation for multihuman multirobot teams. IEEE Transactions on Human-Machine Systems (2024)
8. Mina, T., Kannan, S.S., Jo, W., Min, B.C.: Adaptive workload allocation for multi-human multi-robot teams for independent and homogeneous tasks. IEEE Access **8**, 152697–152712 (2020)
9. Pfleging, B., Fekety, D.K., Schmidt, A., Kun, A.L.: A model relating pupil diameter to mental workload and lighting conditions. In: Proceedings of the 2016 CHI conference on human factors in computing systems. pp. 5776–5788 (2016)
10. Pignoni, G., Komandur, S., Volden, F.: Accounting for effects of variation in luminance in pupillometry for field measurements of cognitive workload. IEEE Sensors Journal **21**(5), 6393–6400 (2020)
11. Scheutz, M., Williams, T., Krause, E., Oosterveld, B., Sarathy, V., Frasca, T.: An overview of the distributed integrated cognition affect and reflection diarc architecture. Cognitive architectures pp. 165–193 (2019)
12. Upasani, S., Srinivasan, D., Zhu, Q., Du, J., Leonessa, A.: Eye-tracking in physical human–robot interaction: Mental workload and performance prediction. Human factors **66**(8), 2104–2119 (2024)
13. Upasani, S.A.: Characterizing mental workload in physical human-robot interaction using eye-tracking measures (2023)
14. Yang, J., Barragan, J.A., Farrow, J.M., Sundaram, C.P., Wachs, J.P., Yu, D.: An adaptive human-robotic interaction architecture for augmenting surgery performance using real-time workload sensing—demonstration of a semi-autonomous suction tool. Human factors **66**(4), 1081–1102 (2024)
15. Zhou, T., Cha, J.S., Gonzalez, G., Wachs, J.P., Sundaram, C.P., Yu, D.: Multimodal physiological signals for workload prediction in robot-assisted surgery. ACM Transactions on Human-Robot Interaction (THRI) **9**(2), 1–26 (2020)