Using Human Workload to Adjust Agent Explanations in Human-agent Teamwork

Zhiqiang Lei
Supervisor(s): Ruben Verhagen, Myrthe Tielman
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Artificial intelligence systems assist humans in more and more cases. However, such systems' lack of explainability could lead to a bad performance of the teamwork, as humans might not cooperate with or trust systems with black-box algorithms opaque to them. This research attempts to improve the explainability of artificial intelligence systems by proposing a framework which models human workload in a value and tailors explanations to this value. Such explanations could provide agents' confidence, causes of making decisions and counterfactual parts to support their suggestions and are adjusted according to agents' knowledge of humans. Results show that adjusted explanations could improve participants' subjective trust in agents and make participants' take more suggestions, while no impact on collaboration fluency or teamwork performance is found.

1 Introduction

Humans and autonomous artificial intelligence (AI) systems are working more and more together in human-agent teams to conduct tasks like illness diagnosis [22] and music composition [24]. The success of such teams is dependent on multiple aspects, such as trust and explainability, and the absence of such factors might lead to a drop of team performance. For instance, if a human could not understand what drives agent's decision, he or she might not cooperate with the agent [18]. Hence, providing explanations for AI systems to improve the efficiency of human-system interaction becomes increasingly significant.

Explainable AI (XAI) refers to AI techniques whose causes for their decisions are understandable to humans, thus making AI systems more interpretable and explainable. In [1], the author states that most recent studies focus on Context-aware explanation method in which agents consider the context when choosing the explanation. On the other hand, User-aware systems are usually ignored, as they only take up 10 percent among all research studied in [1]. User-aware systems are useful because explanations produced by such systems are adjusted or tailored to the user knowledge grasped by the system, making the explanations personalized for different users. In [18], the author concludes that an intelligent agent is able to tailor the explanation to the human observer if it has a model of the human explainee. Though some works [2, 3] prove this by creating user models with different user characteristics, it could be found that models of human workload have not been used to tailor explanations. Human workload is a part of human information which could reveal a human's working status by analysing his or her working history and emotional state. Recent works model human workload mainly by two factors: cognitive load and affective load [7, 6].

This research focuses on how to model and use human workload to tailor explanations including choosing suitable modelling techniques and model usages. The research also examines how the adjusted explanations would affect the team performance, and points out issues found during the experiment to inspire future research directions.

1.1 Research Questions

The main research question is defined as: **How can an agent model and use human workload to tailor explanations?**

To help define the research work clearly, the following sub-research questions are defined:

- Why it is important to model and use human workload?

- Which metrics could be used when modelling the human workload?
- How can the workload model be used when tailoring explanations?
- How to judge whether the tailored explanations are positive for teamwork efficiency?
- What needs to care about when collecting and using the human workload?

The first sub-question refers to the motivation of the research, and the following two questions are about methods used in the research. Last two questions refer to the evaluation metrics and ethical issues.

1.2 Structure

The paper is structured in the following format. In section 2, more background information, including related terminology and works, would be given. Section 3 would talk about the methodology used in the research, consisting of user modelling, explanation tailoring and evaluating methods. Section 4 would give the results of the experiments and statistical and probabilistic analysis. In section 5, the ethical aspects of the research and discuss the reproducibility of methods would be examined. Section 6 would be about the interpretation of the result, reflection on limitations and suggestions for future work. Last sections would conclude the paper.

2 Background and Related Work

2.1 HAT and User-aware XAI

Human-agent teams (HATs) in which humans and AI systems could work together are appearing more in the public eye [26]. Such AI systems are able to observe and record the environment, judge what they have sensed and then make decisions to achieve the predefined goal based on environmental factors. Due to solid memory and the ability of fast computing, agents could perform better than humans with quick and reasonable decisions. However, agents are struggling with emotional issues and unexpected tasks [25], while humans are good at dealing with such situations. Hence, HATs can perform well in different tasks, combining characteristics of both humans and agents.

Team performance of HAT could be affected by multiple factors, such as transparency of AI decision algorithm, explainability and trust. Negative factors could lead to a drop in team performance [11]. Thus, it is necessary to let humans learn about what shapes AI decisions and AI systems are obliged to provide suitable explanations to humans.

Explainable AI (XAI) refers to techniques which attempt to make humans understand AI systems. [5]. One domain of XAI is user-aware XAI, which refers to techniques leveraging human characteristics to tailor explanations [1]. That is, instead of using hard-coded explanations, the system provides specific explanations for humans by tailoring to different factors. The system builds a user model for a specific human, updates the model according to observation and then adjusts the explanation based on the model. A classic case of such systems is EDGE system, which adjusts explanations based on humans' knowledge and their level of expertise [18]. During the dialogue, both factors would be updated based on the information provided by a human, and the agent then outputs relative explanations to make the dialogue more natural. Similarly, BLAH system proposed by Weiner, also has a user model for incremental explanation [27].

2.2 Explanation

In philosophy, an explanation is regarded as both a process and a product [17]. In [15], an act of explaining is stated as someone who possesses causal history of some event attempts to deliver it to someone else. An explanation is considered as "an assignment of causal responsibility" in [12].

In the area of XAI, much work employs causal attribution part of the explanation, which is a process of determining the causes of an event [18]. This is because in most cases, causes are well comprehended and accessible easily by the underlying models. Contrastive explanation shows the cause of counterfactual contrast case that did not happen, instead of the cause of then event itself. As stated in [16] , it is usually easier to answer a contrastive question than to give a complete causal attribution because it is not necessary for one to fully understand all of the causes of the fact. Usually it is enough to know the difference between two cases. Contrastive explanation is important from the view of XAI.

2.3 Human Workload

Various methods have been proposed to measure and model human workload. For measurement, former study shows that it is feasible to measure the team members' workload individually by technology [14], and such measurement could be taken subjectively [23] and by task characteristics. Considering experiments would be taken in a simulated environment, where humans would operate tasks and cooperate with agents in a system, measurement then is not regarded as an important issue because it is only possible for the agents to measure human workworlad by observing his or her behaviour. Instead, workload modelling is considered more important.

Multiple frameworks have been published to model human workload. In [21], authors propose a Cognitive Load and Emotional State (CLES) model to classify and assess human workload more easily. This model maps measured values to cognitive state and emotional state. In the framework, Cognitive Task Load (CTL) consists of three factors, Time Occupied(TO), Level of Information Processing (LIP) and Task-Set Switches (TSS). Time Occupied(TO) describes the proportion of a time when the human is conducting tasks. The second factor, Level of Information Processing (LIP), estimates the mental activity complexity. Task-Set Switches (TSS) show the load of switching from one task to another. Emotional State (ES) works as a complementary part.

In [6], another workload model is proposed. In this research, authors model the workload with Cognitive Load and Affective Load. Cognitive Load is modelled by the same strategy as in [21], and Affective Load consists of Expected Cognitive Load (ECL) and Severity Level of the Current task (SLT). Expected Cognitive Load (ECL) is dependent on the future tasks that humans would take, and Severity Level of Current task (SLT) shows the affective response to the current task.

3 Methodology

This section provides details about the research approach, implementation and how the experiment is conducted.

3.1 Design

Between subject design is employed in this research. The independent variable is whether explanations are tailored to user characteristics or not, and the dependent variables refer to objective scores recorded in the experiment and subjective scores filled in the questionnaire by participants.

3.2 Participants

Participants in this research are humans who are invited to conduct the experiment to control the human agent, and are requested to fill in the questionnaire. In total, 30 participants are recruited by e-mail or oral invitations. Participants are divided into two groups evenly and randomly, the workload group and baseline group according to the types of agents assigned to cooperate with them. Some personal information such as gender and age is collected with the permission of participants. Within 30 participants, 5 of them are female and 25 are male. 21 participants are in the 18-24 years old age group, while 6 are in the 25-34 years old age group. 2 participants are in the 35-44 years old age group and only 1 participant is in the 45-54 years old age group. About their education, 18 of them have achieved some college credits but have not graduated yet, and 10 of them have a bachelor degree. 2 participants have a master degree. About gaming experience, 3 participants do not play video or computer games at all, and 3 participants have a little gaming experience. 8 participants have a moderate amount gaming experience and 6 of them have a normal amount gaming experience. The numbers of participants who have a considerable amount and a lot gaming experience are 6 and 4 respectively.

3.3 Environment

The experiment of this research is conducted in a simulated rescue environment using MATRX framework.

The main task of the experiment is to let humans and robots work together to rescue simulated victims. There are two kinds of victims, critical victims (marked with red colour) and mild victims (marked with yellow colour). Critical victims could only be carried to the drop point by a robot agent and human agent together and rescuing one critical victim gains 6 points for the team, while mild victims could be carried by either a robot agent or a human and rescuing one mild victim brings 3 points. In the environment, there are 4 critical victims and 4 mild victims to be rescued, and they are allocated in some of the 12 blocks. Note that non-player characters marked with green colour are normal ones, which should not be rescued.

To increase the difficulty of finishing the task, there are some obstacles set on the road or on the entry to some blocks. Tree obstacles could only be removed by a robot agent, rock obstacles could be only be removed by a human and robot agent together, and earth obstacles could be removed by either human or robot or together. Some road sections are covered with water. Moving on these water sections would reduce the forward speed of humans and robots. Time limitation of the experiment is set to 8 minutes. Figure 1 presents how the environment looks like with victims' and obstacles' positions marked.

Figure 1: Environment

3.4 Agent Implementing

3.4.1 Human Agent And Baseline Agent

Agents play an important role in this research. As mentioned in the last subsection, the experiment is conducted in a simulated environment, which means human participants do not involve the real rescue task, and instead, they control the human agent to perform the task. Participants could control the human agent by using specified keys. For example, movements of human agents are manipulated by pressing corresponding arrow keys on the keyboard. To reply to the suggestion made by the robot or request help, participants could press buttons in chat-box or directly type it (however the robot could only understand some fixed sentences). Figure 2 is an example of chat-box.

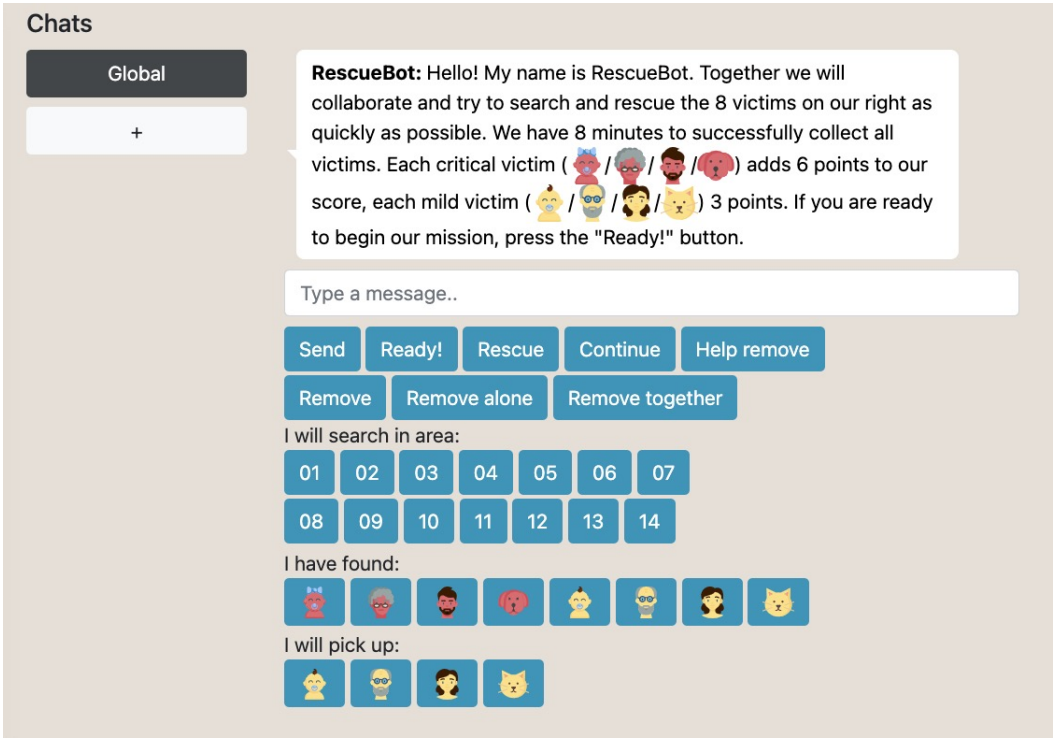

Figure 2: Chat-box

The baseline agent has the basic features of a robot in the rescue scenario. The algorithm for the baseline agent to make decisions is fixed, meaning that the action of the agent is predictive. For instance, after searching a block without finding a victim, or after delivering a victim to the drop point, the agent would go to the nearest unsearched block. Whenever a victim is found, the baseline agent would first ask humans whether to rescue or not. After receiving the decision from humans, the baseline agent would follow this decision definitely. Along with the inquiry are the suggestion and explanation from the robot. Suggestions refer to the action that agent advises human to take, and explanations give the reason to support the suggestion. In this research, there are three types of explanations:

- **Confidence**, simply providing a confidence estimate associated with a certain sugges-

tion, action, etc.

- **Feature explanation**, the reason for the suggestion or what contributed the most to the suggestion.
- **Counterfactual**, what would happen in the alternative suggestion.

When providing the suggestion, the agent randomly append different combinations of explanations to the suggestion. When the found victim is a critical one, the agent would not provide a counterfactual explanation, because there is no counterfactual explanation for this situation.

3.4.2 Workload Agent

The workload agent is generally the same as the baseline agent, except for the ability to tailor explanations to human workload. That is, the workload agent is able to observe and record human workload, then choose a suitable explanation. In this research, human workload is modelled using a semblable structure with [6] framework, *i.e.* Cognitive Load (CL) and Affective Load (AL), but with some adjustments for this scenario.

Cognitive Load CL model is based on Neerincx model [20]. In this model, CL is dependent on three factors: Level of Information Processing (LIP), Time Occupied(TO), and Task Set Switching(TSS). Since this research's goal is to tailor explanations and tailored explanations could affect the Level of Information Processing factor to some extend, this factor will not be considered in the workload model. Thus, Cognitive Load consists of two factors:

- **Time Occupied (TO)**. TO shows how long a human works in a time frame. In this research, the agent could only observe the human action when it is close to a human, resulting in a discontiguous record of human work time. In this case, TO is computed as, the accumulation of human work time divided by the duration of a time frame. A time frame is set to 1 minute, and the accumulation of human work time will reset to 0 every minute.
- **Task Set Switching (TSS)**. TSS factor considers the load of transferring from one task to another. In this research, this factor is computed as the number of tasks divided by the time from a time frame to the start.

Equation below gives how to compute CL at time t:

$$CL(t) = \alpha * TO(t) + (1 - \alpha) * TSS(t)$$

Note that the robot agent could only observe human actions when robot agent is close enough to the human agent, resulting in an inaccurate record of human work time. Thus, α is set to 0.2 to reduce this influence.

Affective Load Usually Affective Load is measured physiologically by human's heart rate or facial expression. In this research physiological measurement is not included, and Affective Load consists of two factors:

- **Severity Level of Current Task (SLT)**. To determine the Severity Level of Current Task, all operations that humans would take during the experiment are divided into three levels, with each level assigning a severity level. Operations are divided into three types: Move, Remove Object and Rescue, with severity levels of 1, 2 and 3 respectively. Severity values for each operation is computed as its severity level divided by 3. Table 1 gives the severity values and corresponding tasks.

Task	Severity Value
MoveSouthEast, MoveWest, MoveNorthEast, MoveSouth, Move, MoveEast, MoveNorth, MoveNorthWest	1/3
RemoveObject , RemoveObjectTogether	2/3
Drop,DropObject, DropObjectTogether, GrabObject, CarryObject, CarryObjectTogether	3/3

Table 1: Severity Values

- **Time Pressure (TP)**. Different from models in [6] and [21], Time Pressure (TP) is considered in this research. Formal study has included TP in affective load when designing human workload model [19]. TP is considered to increase linearly as time passing, and is reduced by 0.1 whenever a critical victim is rescued successfully.

Equation below gives how to compute AL at time t:

$$AL(t) = \beta * SLT(t) + (1 - \beta) * TP(t)$$

In this research, β is set to 0.5.

Workload Final workload value at time t is computed as:

$$Workload(t) = \gamma * CL(t) + (1 - \gamma) * AL(t)$$

Here γ is set to 0.5. All four factors are numeric values at interval [0, 1], so the workload value is also a numeric value at [0, 1].

3.4.3 Tailoring Explanations

Since human workload is expressed as a number value at range [0,1], this range would be split into four intervals to be corresponding to the four types of explanations. According to the research conducted in [4], increase of human workload could result in the decrease of attention percent and increase of reaction time. Inspired by this study, the strategy of tailoring explanations is to provide less explanations when workload is higher. Confidence explanations are considered more important than other two types of explanations, as they tell humans how confident the agent is about its suggestion, which is informative for humans to make a decision. Feature explanations could show the reason contributed to the suggestion, and are considered more important than counterfactual explanations. The detailed tailoring strategy is:

- **[0.00, 0.25]**: Human workload is low, hence it is feasible to provide full explanations.

- **(0.25, 0.50]**: Human workload is relatively low. **Confidence** and **feature** could be combined with suggestion.
- **(0.50, 0.75]**: Human workload is relatively high. **Suggestion** and **confidence** are provided.
- **[0.75, 1.00]**:Human workload is high. Agent only provides **suggestion**.

Note that counterfactual explanations do not exist whenever a critical victim is found, so tailoring strategy for this situation is a bit different: when workload value is lower than 0.50, the agent provides full explanations.

3.5 Measures

This subsection describes dependent variables in this research, *i.e.* the data recorded from the experiment that might be influenced by the independent variable. Dependent variables are measured by two ways: data recorded by the system and data from questionnaires.

Data recorded by the system is objective. For this kind of data, human actions, task completion, scores, ignored suggestions and average human workload are collected.

Data collected from the questionnaire shows the direct feelings of participants and how they judge the task process and result, which consists of subjective human workload, subjective trust, collaboration fluency and explanation satisfaction. To measure subjective human workload, six subjective subscales introduced in NASA-TLX[8] are used. This measurement results on a subjective workload value scaled of [0,100]. 8 questions are set to measure subjective trust, including confidence, predictability of the agent action, reliability, safety, efficiency, warriness, performance and likeability. Another 8 questions are used to mensurate explanation satisfaction. Questions measuring subjective trust and explanation satisfaction are inspired by [10], and scales are [1,5]. 19 questions are used to measure collaboration fluency, which are part of subjective fluency metric scales in [9]. The scale of collaboration fluency is [1,7].

3.6 Procedure

This subsection shows the procedure of a complete experiment from the point of view of a participant. The procedure could be split into four phases:

1. **Preliminary work**. First the moderator of experiment would introduce the basic information of the research and how the experiment is conducted. A consent form is given to each participant to show the purpose of the research and how their data would be preserved and used. All participants that would like to continue next steps should sign on the consent form.
2. **Getting familiar with the environment**. In this procedure, participants will learn how to manipulate the human agent in the environment by following the instructions of a trial agent.
3. **Experiment**. Participants conduct the experiment in this phase. A cheatsheet which indicates the keys for operations and how to remove an obstacle is given.
4. **Questionnaire**. Participants fill in the questionnaire to gather subjective data.

4 Results and Analysis

During conducting the experiment, desired data is collected, which includes objective data recorded by the system and subjective data filled in the questionnaire by participants. For the objective data, scores and numbers of ignored suggestions are interesting. Subjective trust, subjective workload, collaboration fluency and explanation satisfaction are examined as subjective data. Table 2 shows means and standard deviations of these dependent variables, characterized in the workload group and baseline group, according to robot agent type in the experiment.

Variable	**Group**	**mean**	**sd**
score	workload	24.20	6.56
score	baseline	25.00	6.58
suggestions_ignored	workload	0.16	0.13
suggestions_ignored	baseline	0.29	0.16
subjective_trust	workload	3.77	0.38
subjective_trust	baseline	3.46	0.42
subjective_workload	workload	54.17	7.80
subjective_workload	baseline	48.95	13.27
collaboration_fluency	workload	5.39	0.55
collaboration_fluency	baseline	4.94	0.74
explanation_satisfaction	workload	3.85	0.32
explanation_satisfaction	baseline	3.73	0.53

Table 2: Means And Standard Deviations

To check whether there are statistically significant differences between variables of two group, an *independent-samples t-test* is employed. Before conducting the *t-test*, six assumptions need to be checked:

1. Dependent variables are measured on a continuous scale. Dependent variables in this research meet this assumption.

2. Independent variable consists of two independent groups. Two groups are: the workload group and baseline group.

3. Independence of observations. There is no relationship between observations of two groups.

4. There is no significant outlier. This assumption is checked.

5. Dependent Variables should be normally distributed approximately. To check this assumption, *Kolmogorov-Smirnov test* is used. Table 3 shows the variables and their *p-values* of *Kolmogorov-Smirnov test*. A *p-value* greater than 0.05 reveals that data is assumed to be normally distributed. According to this table, only *p-values* of ignored suggestions and collaboration fluency of workload group are lower than 0.05.

6. Homogeneity of variances. For the variables whose data in both groups follow an approximate normal distribution, *Bartlett's test* is employed. For other variables, *Levene's test* is used. All tests show that, for each variable, population variances are same for each group.

Variable	**Group**	**p-value**
score	workload	0.89
score	baseline	0.12
suggestions_ignored	workload	0.04
suggestions_ignored	baseline	0.53
subjective_trust	workload	0.23
subjective_trust	baseline	0.78
subjective_workload	workload	0.72
subjective_workload	baseline	0.93
collaboration_fluency	workload	0.03
collaboration_fluency	baseline	0.54
explanation_satisfaction	workload	0.06
explanation_satisfaction	baseline	0.57

Table 3: Normal Distribution Check

For the variables whose data approximately follow a normal distribution, *independent-samples t-test* is used. For other variables, *i.e.* ignored suggestions and collaboration fluency, *Wilcoxon test* is used. Table 4 presents *p-values* and *mean differences* of these tests. A positive *mean difference* in the table shows that the mean of variable in workload group is higher than the baseline group.

Variable	p-value	mean difference
score	0.74	-0.80
suggestions_ignored	0.02	-0.14
subjective_trust	0.04	0.31
subjective_workload	0.20	5.22
collaboration_fluency	0.07	5.45
explanation satisfaction	0.54	0.12

Table 4: P-values And Mean Differences

In the table, *p-values* of ignored suggestions and collaboration fluency are lower than 0.05, meaning that there are statistically significant differences in the mean scores of these two variables between two groups. It could also be observed that the workload group has higher means of subjective trust, subjective workload, collaboration fluency and explanation satisfaction than the baseline group.

Additionally, a correlation test is done between subjective workload and objective workload of the workload group, resulting in a correlation coefficient of -0.1. This means two variables are independent approximately.

5 Responsible Research

During the research, the most important ethical issue is considered as participants' privacy. This is handled by informing participants how their data would be used and kept. And participants have right to quit the experiment at any moment if they would like. Through a reflection after the experiment, two other issues are detected. First, negative effects of the

experiment should be examined, considering the experiment is conducted in a game environment with a time limit. A study[13] has shown that some video games could cause panic, so participants should be informed of this risk. Second, the experiment uses a simulated rescue scenario and results can not reveal what could happen in the real world.

Detailed information on how agents are implemented, how to set up the experiment and result measurement is given in Section 3. Other information such as participants information could also be found in Section 3. Such information is helpful for reproduction of this research.

6 Discussion

This section interprets the results, talks about the limitations of the research, and fives possible future work directions.

6.1 Interpretation of Results

Statistics in Section 4 show that there exists significant difference between ignored suggestions in the workload group and baseline group. With a positive mean difference, it could be presumed that participants ignore less suggestions with a workload agent. Statistics also show that notable difference is found between subjective trust values in the workload group and baseline group, and the mean of the workload group is higher than the baseline group. Besides ignored suggestions and subjective trust, no statistically significant difference between means of other four variables is detected. These variables are: score, subjective workload, collaboration fluency and explanation satisfactory.

It can be regarded that participants tend to ignore less suggestions provided by the workload agent, and they tend to give higher subjective trust values when filling out questionnaires. This may because, explanations provided by the workload agent are in conformity with a fixed principle, *i.e.* less explanations are provided when objective workload is higher, and the baseline agent gives different combinations of explanations randomly. Interestingly, no significant difference is detected between explanation satisfaction, meaning that participants do not feel more or less satisfactory with explanations tailored by the workload agent. The reason of this could be defects of tailoring strategy. Additionally, they do not experience an improvement or drop of the collaboration fluency and their subjective workload values do not show a rise or drop. This might because explanations do not take a huge share in the collaboration, or tailored explanations are not good enough to shape the teamwork . Participants in the workload group do not achieve higher or lower scores than baseline group, either. For example, for the Severity Level of Current Task factor in Affective Load, tasks of removing obstacles are given a severity value of 2/3, while in the real world participants could stay idle for a duration as it would take some time to complete the remove.

Besides, a correlation coefficient of -0.1 shows there is almost no correlation between human objective workload and subjective workload, which could also reveal deficiencies of workload modelling. Such deficiencies could be due to the inconsistency of the real world to the simulated environment. For example, for the Severity Level of Current Task factor in Affective Load, tasks of removing obstacles are given a severity value of 2/3, while in the real world participants could stay idle for a duration as it would take some time to complete the remove.

6.2 Limitations

In the research, a human workload model is built by the robot agent to record the workload of a human agent and then choose suitable explanations to give reasons how the suggestion of the robot agent is made. The experiment is done in a simulated rescue scenario with human participants operating the human agent. Measurements are based on the objective values recorded by system and subjective values from questionnaires filled in by participants.

By checking the experiment results and reviewing on experiment setups, some limitations could be identified.

Difference between simulated environment and the real world. The experiment is conducted with participants operating human agents in a simulated environment. Workload is actually modelled based on actions of the human agent manipulated by participants, hence may not reveal the true feelings of participants .

Whether participants check explanations carefully. As the experiment is fast-paced (time limitation is 8 minutes), and according to the response data, many participants have a high mental load when conducting it. It is hard to tell whether all participants read the explanations carefully.

6.3 Future Work

Considering the limitations talked in the last part, some future work directions and improvements could be determined:

1. Improvement of workload model. In this study human workload is computed based on agents' observation. As mentioned before, this may not reveal participants' true workload correctly . Hence, more accurate measurements could be employed, such as recording human's physiological data.
2. Examine whether participants check explanations carefully or not.
3. More dynamic method of tailoring explanations. Explanations could be tailored according to the distribution of Cognitive Load, Affective Load or other factors. More combinations of explanations could be explored.

7 Conclusion

This research focuses on tailoring explanations to human workload and proposes a framework which models human workload with two factors and chooses corresponding explanations by mapping workload value into four scales. Results have shown that participants ignore less suggestions and trust agents more with tailored explanations, while no evidence shows that collaboration fluency and performance of human-agent teams could be influenced by explanations tailored to human workload. It is also worth mentioning that there is no correlation found between objective human workload and subjective human workload, which might be due to the difference between simulated environment and the real world. A future study could consider a new form of human workload model which is more consistent between system measurement and humans' real feelings. Besides, explanations could be tailored more flexibly.

References

[1] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS â19, page 1078â1088, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

[2] Alison Cawsey. User modelling in interactive explanations. *User Modeling and User-Adapted Interaction*, 3(3):221–247, 1993.

[3] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Explanation generation as model reconciliation in multi-model planning. *CoRR*, abs/1701.08317, 2017.

[4] Mehran Ghalenoei, Seyed Bagher Mortazavi, Adel Mazloumi, and Amir H. Pakpour. Impact of workload on cognitive performance of control room operators. *Cognition, Technology Work*, 24(1):195–207, 2022.

[5] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. Darpa's explainable ai (xai) program: A retrospective. *Applied AI Letters*, 2(4):e61, 2021.

[6] Maaike Harbers, Reyhan Aydogan, Catholijn Jonker, and Mark Neerincx. Sharing information in teams: Giving up privacy or compromising on team performance? volume 1, 05 2014.

[7] Maaike Harbers and Mark A. Neerincx. Value sensitive design of a virtual assistant for workload harmonization in teams. *Cognition, Technology Work*, 19:329–343, 2017.

[8] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[9] Guy Hoffman. Evaluating fluency in humanârobot collaboration. *IEEE Transactions on Human-Machine Systems*, PP:1–10, 04 2019.

[10] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects, 2018.

[11] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna van Riemsdijk, and Maarten Sierhuis. Coactive design: Designing support for interdependence in joint activity. *J. Hum.-Robot Interact.*, 3(1):43â69, feb 2014.

[12] John R. Josephson and Susan G. Josephson. *Abductive Inference: Computation, Philosophy, Technology*. Cambridge, England: Cambridge University Press, 1994.

[13] Judith Lauter, Elizabeth Mathukutty, and Brandon Scott. How can a video game cause panic attacks? i. effects of an auditory stressor on the human brainstem. *The Journal of the Acoustical Society of America*, 126:2204, 01 2009.

[14] S. Levin, D. Aronsky, R. Hemphill, J. Han, J. Slagle, and D. J. France. Shifting toward balance: measuring the distribution of workload among emergency physician teams. *Ann Emerg Med*, 50(4):419–23, 2007.

[15] David Lewis. Causal explanation. In David Lewis, editor, *Philosophical Papers Vol. Ii*, pages 214–240. Oxford University Press, 1986.

[16] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247â266, 1990.

[17] Tania Lombrozo. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470, 2006.

[18] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[19] Diane Nahl. Affective and cognitive information behavior: Interaction effects in internet use. *Proceedings of the American Society for Information Science and Technology*, 42, 10 2006.

[20] Mark Neerincx. Cognitive task load analysis: allocating tasks and designing support. *Handbook of Cognitive Task Design. Chapter 13. Mahwah, NJ: Lawrence Erlbaum Associates*, pages 283–305, 01 2003.

[21] Mark A. Neerincx, Maaike Harbers, Dustin Lim, and Veerle van der Tas. Automatic feedback on cognitive load and emotional state of traffic controllers. In Don Harris, editor, *Engineering Psychology and Cognitive Ergonomics*, pages 42–49. Springer International Publishing.

[22] B. N. Patel, L. Rosenberg, G. Willcox, D. Baltaxe, M. Lyons, J. Irvin, P. Rajpurkar, T. Amrhein, R. Gupta, S. Halabi, C. Langlotz, E. Lo, J. Mammarappallil, A. J. Mariano, G. Riley, J. Seekins, L. Shen, E. Zucker, and M. Lungren. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit Med*, 2:111, 2019.

[23] Gary B. Reid and Thomas E. Nygren. The subjective workload assessment technique: A scaling procedure for measuring mental workload. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 185–218. North-Holland, 1988.

[24] Adam Roberts, Jesse Engel, Yotam Mann, Jon Gillick, Claire Kayacik, Signe NÃžrly, Monica Dinculescu, Carey Radebaugh, Curtis Hawthorne, and Douglas Eck. Magenta studio: Augmenting creativity with deep learning in ableton live. In *Proceedings of the International Workshop on Musical Metacreation (MUME)*, 2019.

[25] Filippo Santoni de Sio and Jeroen van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5, 2018.

[26] Jurriaan van Diggelen, Jonathan Barnhoorn, Marieke M. M. Peeters, Wessel van Staal, M. L. van Stolk, Bob van der Vecht, Jasper van der Waa, and Jan Maarten Schraagen. Pluggable social artificial intelligence for enabling human-agent teaming. *CoRR*, abs/1909.04492, 2019.

[27] J.L. Weiner. Blah, a system which explains its reasoning. *Artificial Intelligence*, 15(1):19–48, 1980.