The Impact of Tailoring Agent Explanations According to Human Performance on Human-AI Teamwork

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

June 20, 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

The Impact of Tailoring Agent Explanations According to Human Performance on Human-AI Teamwork

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

June 20, 2022

Abstract

Nowadays, artificial intelligence (AI) systems are embedded in many aspects of our lives more than ever before. Autonomous AI systems (agents) are aiding people in mundane daily tasks, even outperforming humans in several cases. However, agents still depend on humans in unexpected circumstances. Thus, the main goal of these agents has transformed from becoming independent to interdependent systems, collaborating with humans. This collaboration is far from perfect and could be improved in several aspects. Communication is crucial for flawless collaboration and its key aspect is explainability. This paper studies the impact of tailoring explanations according to human performance in a well-defined collaborative human-agent teaming (HAT) urban search-and-rescue (USAR) task environment. A controlled experiment was conducted in a between-subject manner, with two different agent implementations, where it was hypothesised that when an agent provides explanations tailored to human performance, the collaborative performance, the trust towards the agent and the individual satisfaction of the human would increase. Results of the experiment confirmed that this is indeed the case for explanation satisfaction, however, not necessarily for trust and performance metrics. The conclusions also included that the tailoring resulted in a decreased collaborative performance. The research contributes to the bigger picture of how tailoring explanations to various factors, would have an impact on the overall collaborative performance and systematic actualisation of HAT.

1 Introduction

The rapid developments in artificial intelligence (AI) technology are allowing autonomous artificially intelligent systems (agents) to become a fundamental part of our lives. From virtual personal assistants to unmanned aerial vehicles, research and practical experiences in the field have proven the benefits of agents in complex, quantitative and defined circumstances [1, 2]. In fact, as they improve, the general consensus has become that they will completely overtake tasks of humans and become **independent** of any human aid [2]. However, this is not the case; further developments have shown that a world where agents completely take over the tasks of a human is far in the future, and maybe not even feasible [3, 4, 5].

Despite agents being used in many sectors, and outperforming humans in certain areas, they have been observed struggling with tasks involving emotions, moral decisions and/or unexpected events, even in the most developed areas of AI [6, 7]. Human support and supervision are deemed necessary, especially in tasks with high consequences. Considering it all, the focus of the research field has shifted from creating independent agents to adapting them to become **interdependent**, by **collaborating in human-agent teaming (HAT)** [4, 5, 8, 9].

In order to develop an *'artificial'* intelligence, human nature has to be examined [10]. Creating a seamless and authentic experience for humans is the biggest challenge while developing human-agent collaborative systems, and the most effective aspect of this experience is also the most difficult for an agent, which is **communication** [9, 11, 12, 13, 14].

While communicating to collaborate with someone, explanations about relevant information behind any fact or advice are one of the most influential aspects [14]. However, this is quite a challenge for agents, "because people assign human-like traits to artificial agents, people will expect explanations using the same conceptual framework used to explain human behaviour" [11, p. 2]. As human communication is a complicated framework depending on many factors including context, people involved, emotions etc, these dependencies should also be represented in artificial explanations. Thus, the shift has caused an expansion in the research field of AI. At the intersection of Social Science, Human-Computer Interaction and AI itself, a new field has emerged, called **eXplainable AI (XAI)**[1] [11].

The broad and complex nature of explainability in human sociology also comes into sight in the domain of AI. Its main goal is to make AI systems more understandable to humans, making them **personalised and user-aware** [15, 16]. There are endless factors that influence the adaptation of the communicated explanations of a human being that needs to be identified and artificially replicated. To understand their impact on collaboration, they need to be studied individually.

Currently, the research on XAI includes several studies about user-aware and context-aware solutions. However, there is a clear research gap in tailoring explanations according to specific human (user) related factors. Of all the factors, this paper specifically focuses on **performance** as it is one of the most significant. It directly gets influenced by and has the potential to enhance individual performance, which ultimately affects collaborative performance. Thus, it is crucial to allow the agent to consider the performance of its human teammate while providing explanations and examining its impact on HAT. The significance of the performance factor is further described in Section 2.

1.1 Research Question and Hypothesis

Within TU Delft[2], the dedicated AI*MAN Lab[3] and research group have been working on transparent XAI systems, where, for example, explanations are classified into several types with a two-dimensional model in the work of Verhagen [17]. The research topic of this paper, "User-aware eXplainable AI for improving human-AI teamwork" is supervised by PhD Candidate Ruben Verhagen and Responsible Professor Myrthe Tielman.

This research paper aims to answer; **"How can an agent model and use *human performance* to tailor explanations?"**. This question will be analysed under several sub-questions, in a step-by-step manner:

[1] ibm.com/watson/explainable-ai
[2] tudelft.nl
[3] tudelft.nl/en/ai/aiman-lab

1. What metrics can be used to evaluate and model human performance in a given collaborative task scenario?
2. How can explanations adapt according to a human performance model?
3. How can two models of explanations be compared in an experimental setup?
4. How can human performance and satisfaction be evaluated according to implemented adaptive agent explanations?
5. How does collaborative performance, human trust and satisfaction towards the agent change according to implemented tailored agent explanations?

Common sense says that collaboration generally increases efficiency in human teams, as depicted in the famous saying; *"two heads are better than one"*. Thus, it is also reasonable to assume that collaboration with AI agents could even surpass that increase, with the right amount of aid from each team member, through efficient communication. In other words, the hypothesis of this paper argues that an agent with tailored explanations according to human performance could increase understanding and improve communication, which could result in better collaborative performance, increased human satisfaction and trust towards the agent.

In the big picture, the conclusions derived from this research topic could contribute to a world where agents and humans communicate and collaborate efficiently in a variety of applications. The future of XAI lies in the technique of combining all factors into a single adaptive explanation methodology and its evaluation in a variety of settings, possibly with more research on domain-specific improvements.

2 Background and Related Work

As with any type of collaboration, the ideal solution leverages the expertise of each team member and yields the best performance result [18]. This is also the case in HAT. "An effective human-AI team ultimately augments human capabilities and raises performance beyond that of either entity" [1, p. 1]. Thus, in HAT, collaboration and coordination are essential [19]. With many dependencies between the human and the agent, information and thought sharing, together with a reasoning explanation would lead to the most effective communication model, increasing collaborative performance [10, 14, 20, 21, 22].

A successful HAT communication depends on several factors, such as transparency, mutual trust, understandability and explainability. Understanding any action done or decision made by the agent is done through explanations of the reasoning. This allows the human to understand the inner workings of the agent [10, 17, 18, 21, 23, 24].

In contemporary HAT, many of the aforementioned factors are still lacking. This can often be seen in current AI systems with poor explainability, leading to insufficient understanding for humans about the inner logic of agents, resulting in decreased collaborative performance. With insufficient transparency and explainability, all other metrics that allow perfect collaboration, such as trust in the agent, mutual understanding, and overall collaborative performance, also decrease [8, 10, 11, 16, 21, 22].

Avoiding this is only possible by trying to replicate human intelligence and for that, dynamic explainability according to several sub-factors is required. The agent's methodology

must easily adapt to the intended user and use case by correctly identifying and processing these factors. In terms of being user-aware, several sub-factors can be used to tailor explanations, one of which is performance [16, 17, 25].

From common sense, it can be derived that performance is one of the most crucial factors used in human teaming scenarios. For example, while playing basketball, one never explains the best shooter of the team, how to shoot, or does not rely on the worst shooting player to complete a complex move without any explanation. To be able to play as a team, other players should be able to incorporate the worst player into the game plan. Basketball crucially depends on one's shooting capabilities, in addition to several sensing and reasoning attributes. Most importantly, it is a team sport where members must communicate and collaboratively improve on these attributes by learning from each other, which is done by tailoring explanations according to performance.

In the paper called "Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity" [26], Klein et al. implicitly state tailoring explanations to human performance as a challenge. 'To be an effective team player, intelligent agents must be able to adequately model the other participants' intentions and actions vis-à-vis the joint activity's state and evolution-for example, are they having trouble? Are they on a standard path proceeding smoothly? What impasses have arisen? How have others adapted to disruptions to the plan?" [26, p. 92]. This research will also contribute to Klein's second challenge by providing evidence on how performance-tailored explanations affect an agent becoming a 'team player'.

As with many real-life attributes, human performance is also not binary, it consists of a spectrum, and changes according to time, occurrence and severity. Specific performance attributes all correspond to a different game mechanism, where different explanations are communicated. Thus, correctly tailored explanations can have a positive impact on the recipient's individual performance and therefore on the collaborative performance [9, 21]. This research will incorporate all mentioned aspects of the human performance factor into explanations and replicate this scenario in an urban search-and-rescue task while observing its influence on collaboration.

3 Methodology

Incorporating all the information discussed in Section 2, two agents, demonstrating default (baseline) and tailored explanations were implemented. These two agents were teamed-up with human participants in a collaborative task environment, and each experiment was evaluated over several metrics. Comparing these metrics and results provided the answer to the research question at hand.

3.1 Design

The goal of the experiment was to investigate whether explanations tailored to human performance affect collaborative performance, human satisfaction and trust. To achieve this, a controlled experiment was conducted in a between-subject manner, where half of the participants teamed up with the Baseline Agent (described in Section 3.4.2) and the other half teamed up with the user-aware agent with tailored explanations (Tailored Agent, described in Section 3.4.3).

The Baseline agent had a default, generic and sometimes no explanations for the messages it sent to the human, while the other agent included explanations specifically tailored to

human performance.

3.2 Participants

Participants were recruited through the TU Delft bachelor's students network from different majors. The inclusion criteria were: 18-60 years of age, at least a high school degree and sufficient computer literacy to play a game. Exclusion criteria included colour blindness.

A total of 30 participants took part (27 male, 2 female, and 1 preferred not to say) in the experiment. 26 participants were in the 18-24 age range while the other 4 were under 60 and close to the same range. The median score of self-scored computer gaming experience was 4.2 for the tailored agent, and 3.9 for the baseline agent (1, meaning no experience, and 6 meaning a lot of experience) and the median education level was 4.2 for baseline (0, meaning no education experience at all, 4 meaning some college credit/no degree yet and 8 meaning a PhD degree). No data were excluded from the analysis since all experiments were done as planned.

3.3 Environment

The environment used in the experiments was formulated as an urban search-and-rescue game. The game made use of the MATRX Software. The details about the software and specifics of the game mechanics are described below.

3.3.1 MATRX Software

With the increased interest in human-agent teaming research, the need for an easy-to-use research tool for a modifiable HAT environment emerged. For several years, the Blocks World for Teams (BW4T)[4] [27] software for human-agent teaming studies has been used. It is an extension to the classic AI planning problem of Block World (BW) [28], adding an easy-to-configure teaming aspect. However, these tools are often outdated, unreliable and simple, as they allow limited dimensionality in HAT research.

In recent years, a new tool has been developed by TNO[5], named Human-Agent Teaming Rapid Experimentation Software package (MATRX)[6]. This new tool offered more advanced support on HAT scenarios with additional environmental configurations. It allowed researchers to create more personalised and detailed HAT environments to test and study the impact of many variables.

One of its capabilities is to easily create a simulated urban search-and-rescue (USAR) human-agent teaming (HAT) scenario. The USAR-HAT scenario is now used in many HAT studies and it is also the best fit implementation for the experiment of this paper.

3.3.2 Collaborative Urban Search-and-Rescue Task

The best way of demonstrating the effect of factors in HAT is to use a task environment with several interdependent and independent sub-tasks, which USAR easily provides [29].

As a summary of the game mechanics, the human and the agent try to explore several rooms (areas) in a two-dimensional game map. They are tasked to find 4 critically injured and 4 mildly injured victims, searching through those areas and rescuing them to a dedicated

[4] github.com/eishub/BW4T
[5] tno.nl
[6] matrx-software.com

'drop-zone' on the map. As they search for victims in different areas, they encounter some obstacles that they must remove. Different obstacles take different times to remove and some can only be removed together or only by the agent. A stone can be removed alone, but it is faster together, a rock can only be removed together and a tree can only be removed by the agent. This is also the case with the victims, where mildly injured ones could be carried individually, but the critically injured victims need a collaborative effort to be rescued.

The game lasts 8 minutes and works on a point system. Each mildly injured victim that is rescued is worth 3 points and each critically injured victim that is rescued is worth 6 points. The game aims to rescue most victims and gain the most points in time by collaborating with the agent.

The MATRX environment includes a chat feature, which allows communication between the agents. In the USAR task, the agents can inform each other about the area they are going to search, the victims/obstacles they found, if they are going to rescue/remove them, etc. The collaborative tasks force the human and the agent to communicate, where the explanations of the agent shape the response of the human. A screenshot of the game environment together with the chat feature is attached to Appendix A.

3.4 Implementation of the Agents

The game is played in teams of two, a human participant matched with an agent teammate. In the game, the agent leaves all the decisions about removing an obstacle or rescuing a victim, to the human. In the meantime, the agent acts as a decision support agent. It gives suggestions about the decision to be made and leaves the final decision to the human. While giving these suggestions, or taking any action, the agent explains the reasoning behind them. This explanation is the core difference between the two agent implementations.

3.4.1 Pilot Experiment for Explanations

As the agent was designed as a decision support agent, it had to give suggestions for the next move it makes. Since some actions required collaboration, where others did not, the decision was always left to the human, however, the agent had to give a suggested decision, backed by an explanation.

A pilot experiment was necessary to create the explanations used in our experiment. To achieve this, a crowd-sourced survey was used, where participants were asked to decide on the next action of the agent in a given scenario.

The scenarios were created with differing locations, distances from teammates distances from the 'drop-zone' and time left for the game. They were created based on intuition and prior knowledge, and in the implementation, it was made sure that the question metrics were accurate with in-game real-time metrics when the suggestion was used.

In the survey, nine participants were confronted with dilemmas about whether to remove obstacles or save victims, based on several different features such as time left, removal time, distance to the teammate and critical victims rescued. After submitting their decision, participants were also asked about which feature contributed most to their decision.

This experiment allowed us to determine a default reasoning for the baseline agent explanations. There are several types of explanations an agent can use. The categories of explanations used were described in the study of Van Der Waa et al. [30]. We have made use of four combinatory explanation types: suggestion, suggestion + confidence explanation, suggestion + confidence explanation + feature attribution, and suggestion + confidence ex-

planation + feature attribution + contrastive explanations. The collected data was used to create the four types of explanations.

Out of the nine participants, the decision most participants made was transformed into the suggestion for similar scenarios. The number of participants out of the total nine, who agreed on that decision was transformed into the confidence explanation, where higher people meant higher confidence. The feature selected by the most participants was transformed into the feature attribution. Finally, the contrastive explanation was created based on what is the main difference between another scenario where the other option was selected. All explanations included an observation at the beginning and a call for a decision at the end. Below is an example of a full explanation, divided into parts:

1. **Observation:** Found rock blocking area 4.
2. **Suggestion:** I suggest to remove rock instead of continue searching
3. **Confidence Explanation:** : 5/9 rescuers would decide the same,
4. **Feature Attribution:** because we have around 5 minutes left.
5. **Contrastive Explanation:** If we had found less than 2 critical victims, I would have suggested to continue searching.
6. **Call for Decision:** Select your decision using the buttons "Remove" or "Continue"

3.4.2 Baseline Agent Implementation

The agent was programmed to try to win the game by searching all rooms (each time picking the closest one) and rescuing all victims. It made use of the chatting functionality of the environment to inform its teammate about where it is heading, which obstacles/victims it found and if it is rescuing a victim or removing an obstacle.

Whenever the agent faces a decision to be made, the global problem solving team design pattern was used [31]. The agent asked the human to decide its next action after providing its suggestion, by randomly picking one of the four explanation types detailed above. This was the best suitable option to be able to compare it with a tailored version of the agent. While waiting for the decision, the agent did not take any action.

3.4.3 Tailored Agent Implementation

The tailored agent implementation was written on top of the baseline agent. It additionally, kept track of the performance statistics of the human and the game, while also using that data to tailor its explanations. The tailoring procedure was also implemented by the modifications to the existing explanations given by the baseline agent, with one additional message type added.

The overall performance of the team depends on many factors explained above. The 'mission' finishes with a higher success rate when collaboration and communication are done efficiently. The best player in the game, cannot finish it with a high score if he or she does not frequently update the agent and answer its questions in time. That is why messaging was considered a crucial performance aspect of the game. Moreover, the 'chemistry' or coordination between teammates was also another aspect that stood out as a performance factor, where both agents must be efficiently trying to help and trust each other. Finally, the real-time score of the game was the most basic performance statistic. Thus, the factors

used to tailor explanations could be categorised into three types, messaging performance, collaboration performance and score performance.

Messaging Performance For messaging performance, the agent tracked the messages of the human, where two main metrics determined the messaging performance. The first one was implicitly tracking the time it takes the human to search a room. This was the time measured between a 'Searching room X' and a 'Found victim X' or another 'Searching room X' message. This meant that during that time the human is searching the first room he/she mentioned. The agent checked if this time exceed a maximum expected duration, which was calculated through the game logic where in 8 minutes, 2 team members must visit all 14 rooms. If this time has been exceeded, the agent's upcoming explanation was tailored accordingly with an addition of a reminder to message more frequently. The second messaging metric was the time it took the human to rescue a found victim. This was measured as the time between a 'Picking-up victim X' and the next 'Searching room X' message. Using the same game logic from before, a threshold was determined and after it was exceeded the explanation was tailored accordingly. If the thresholds were not exceeded there was a possibility of the human receiving a motivational message, together with a tip about the game.

Collaboration Performance For collaboration performance, two metrics were stored. The first was, again, a messaging metric which tracked the time it took the human to answer the agent's questions. It was crucial for the human to answer the agent's questions as fast as possible, to minimise the idle time of the agent. During the processing of this metric, the agent also tracked the number of times this 'error' of a slow answer was made, however, this metric was reset when the human answered a question quickly. A reminder message got sent after a large threshold was passed, and an additional sentence was added to the next request after a smaller threshold was exceeded. The severity of the sentences changes according to how many times the error was done.

The second metric tracked the time it took for the human to come to the agent's aid when the agent needed help. However, if the human is busy with another task, but promises to come and help the agent, this causes the team to lose significant time. Moreover, this metric also kept track of how many times this 'error' was made, with the same resetting rule. Same as the answering metric, the next time agent asked for help, an additional reminder was sent to the human to be faster this time, again, the severity of this message changed.

Score Performance Each game tick (one-tenth of a second) the agent calculated the difference between percentages of the expected and the real-time score. Out of 36 total points, the team was expected to gain 4.5 points each minute, or 18 points at the 4-minute mark. Through intuition and prior knowledge of the game mechanics, several threshold values were determined to convert this difference value into a four-point performance spectrum (very poor, poor, good, very good). A difference larger than 20% meant that the human was performing very poorly and a difference smaller than 6% meant that the human was performing very good.

This value was used in two ways in the implementation of the agent. The spectrum levels determined the type of explanation the human would receive. To improve the performance of an under-performing human, the agent provided more detailed descriptions. On the contrary, it provided concise explanations for the top-performing human. Also, the human occasionally received a progress message, including the team's predicted score and

an explanation of how and why it was calculated that way together with a motivational sentence. Under-performing humans were more likely to receive this message or one that explains why high-confidence recommendations should not be ignored.

3.5 Measures

The dependent variables of the study include several objective and subjective measures. Objective measures were recorded through the logging feature of MATRX (Log) and the subjective data was collected through a questionnaire that included five different surveys.

To ensure the minimal effect of the confounding factors, some demographic data was collected before each experiment. One of the most important confounding factors was the gaming experience of the participants, which was thought to be able to influence the performance difference between the two experimental groups. To create a similar division, demographic distribution was considered while directing participants to the experiment group. All taken measurements are explained in detail in Table 1.

3.6 Procedure

The experiment underwent an extensive ethical evaluation process and took into account the principles of responsible research described in more detail in Section 6.

It took 25 to 35 minutes for a participant to fully complete the experiment. Before the experiment, the participants were given an introduction and briefed about the purpose and procedure of the experiment. Afterwards, they were given an "Informed Consent Form" and asked to read and sign the form.

Participants were then given a laptop and a mouse by the experiment leader to first access the Qualtrics[7] questionnaire. After filling in the demographics questions, they were assigned to the tailored agent experiment group or the baseline agent experiment group.

Then, the experiment leader gave the participants a cheat sheet, including important reminders about the game mechanics and started the tutorial, which is a practice round, teaching the game mechanics in a hands-on manner. This allowed the participants to be more comfortable and gain experience in the game environment.

After the tutorial game, the participants were reminded that they will have to decide on each action of the agent ('RescueBot'). Then, the experiment leader started the actual game for the participant according to his/her experiment group.

All participants completed one full game of 8 minutes. After finishing the game, the game logs were saved by the experiment leader and the participants were led to the remaining parts of the questionnaire, where subjective questions about collaboration fluency, explanation satisfaction, trust towards the agent and workload were answered, respectively.

4 Results and Analysis

All of the metrics discussed in Section 3.5, were statistically analysed for significant differences between the baseline and tailored agent groups. A Classical Independent Two-Samples T-test (with significance level $\alpha = 0.05$) was performed for any metric that satisfied the assumptions, including independence of the observations, identification of extreme outliers, normality of the data (checked by the Shapiro-Wilk test, significance at $\alpha = 0.05$) and homogeneity of variances (checked by the Levene's test, significance at $\alpha = 0.05$). Moreover,

[7] qualtrics.com

Table 1: Measurements taken during the experiment. Quant = Quantitative Data, Qual = Qualitative Data, Obj. = Objective Metric, Subj = Subjective Metric.

Metric (Source)	Data Type (Range)	Description
Demographic Information (Survey)	Subj./Obj. & Quant./Qual.	A small survey, asking about their gender, age range, education level and gaming experience (ranging from 1 (none at all) to 5 (a lot)).
Ignored Suggestions (Log)	Obj. & Quant. (0-1)	The percentage of ignored suggestions is calculated by dividing the number of ignored suggestions by the total suggestions made.
Subjective Trust (Survey)	Subj. & Quant./Qual. (1-5)	Mean Value of the Trust Scale for XAI Survey [32], measured by a 5-point Likert scale (ranging from I disagree strongly to I agree strongly).
Explanation Satisfaction (Survey)	Subj. & Quant./Qual. (1-5)	Mean Value of the Explanation Satisfaction Survey [32], measured by a 5-point Likert scale.
Completeness (Log)	Obj. & Quant. (0-1)	The completeness percentage of the game was calculated via the number of rescued victims divided by the total victim number.
Score (Log)	Obj. & Quant. (0-36)	The score of the game was calculated by rescued victims, where critically injured ones were worth 6 points and mildly injured ones were worth 3 points.
Agent Moves (Log)	Obj. & Quant.	Total number of moves made by the agent.
Human Moves (Log)	Obj. & Quant.	Total number of moves made by the human.
Agent Idle Time (Log)	Obj. & Quant.	Total game ticks where the agent is idle.
Human Idle Time (Log)	Obj. & Quant.	Total game ticks where the human is idle.
Simultaneous Idle Time (Log)	Obj. & Quant.	Total game ticks where the human and the agent are idle at the same time.
Subjective Workload (Survey)	Subj. & Quant./Qual. (0-100)	Mean Value of the Raw NASA-TLX Survey [33], measured by an adjustable 100-point scale.
Collaboration Fluency (Survey)	Subj. & Quant./Qual. (1-7)	Mean Value of the Collaboration Fluency Survey (20 out of 30 Questions) [34], measured by a 7-point Likert scale. 'Robot Relative Contribution' and 'Individual Measures' metrics were left out since they were not related to the research question and 'Trust in Robot' was left out because it was already measured by a separate variable.

all collected metrics were continuous. Out of all the data, only one sub-metric included an extreme outlier, however, since its exclusion resulted similarly, the data point was kept to preserve the true data. In metrics where normality was not found the Unpaired Two-Samples Wilcoxon test was applied.

First of all, only the gaming experience demographic data were analysed for any significant difference, or any correlation with several other metrics, since the rest was used only to divide participants equally. The gaming experience metric was thought to be a confounding factor, and thus, was further analysed. Wilcoxon test applied to *gaming experience* metric showed that the difference was insignificant with p = 0.6264 (Baseline Median = 4 (Interquartile Range = 2.5), Tailored Median = 5 (IQR = 1.5)).

Next metrics used to determine human trust were analysed. For the *ignored suggestions* metric, the t-test resulted t(28) = -0.101, p= 0.921 indicating no significant difference (Baseline M = 0.294 (SD = 0.163), Tailored M = 0.301 (SD = 0.198)), where, t(28) is shorthand notation for a classical t-statistic that has 28 degrees of freedom, while 'M' and 'SD' are shorthand notations for mean and standard deviation. For the *subjective trust* metric, the t-test resulted t(28) = -1.04, p = 0.307, indicating no significant difference (Baseline M = 3.46 (SD = 0.416), Tailored M = 3.65 (SD = 0.577)).

Regarding satisfaction of the human, the mean *explanation satisfaction* in baseline group was 3.74 (SD = 0.536), whereas the mean in tailored group was 4.23 (SD = 0.36). The t-test showed that the difference was **statistically significant**, t(28) = -0.295 with p = 0.00631, as depicted in Figure 1.

Analysis on objective performance metrics resulted as follows. The median *completeness* in baseline group was 0.75 (IQR = 0.25), whereas the median in tailored group was 0.625 (IQR = 0.125). The Wilcoxon test showed that the **difference was significant** with p = 0.05666. The mean *score* in baseline group was 25 (SD = 6.58), whereas the mean in tailored group was 19.2 (SD = 6.88). The t-test showed that the difference was **statistically significant**, t(28) = 2.36 with p = 0.0255, as depicted in Figure 2. The mean *agent moves* in baseline group was 290 (SD = 84.7), whereas the mean in tailored group was 236 (SD = 46.9). The t-test showed that the difference was **statistically significant**, t(28) = 2.18 with p = 0.0377. Wilcoxon test applied to *human moves* metric showed that the difference was insignificant with p = 0.08139 (Baseline Median = 467 (IQR = 109), Tailored Median = 412 (IQR = 60)). For the *agent idle time* metric, the t-test resulted t(28) = - 0.405, p = 0.688, indicating no significant difference (Baseline M = 1774 (SD = 811), Tailored M = 1878 (SD = 580)). Wilcoxon test applied to *human idle time* metric showed that the difference was insignificant with p = 0.1102 (Baseline Median = 3261 (IQR = 510), Tailored Median = 2756 (IQR = 776)). For the *simultaneous idle time* metric, the t-test resulted t(28) = -0.361, p = 0.721, indicating no significant difference (Baseline M = 1368 (SD = 699), Tailored M = 1454 (SD = 591)).

Finally, subjective performance metrics were analysed. For the *subjective workload* metric, the t-test resulted t(28) = -0.62, p = 0.54, indicating no significant difference (Baseline M = 48.9 (SD = 13.3), Tailored M = 51.6 (SD = 10.1)). For the *collaboration fluency* metric, the t-test resulted t(28) = -0.473, p = 0.64, indicating no significant difference (Baseline M = 4.94 (SD = 0.738), Tailored M = 5.08(SD = 0.847)). Sub-metrics of collaboration fluency were also analysed for spesific significance, where only one of the metrics produced a significant difference between conditions. The median *'improvement'* metric in baseline group was 4.67 (IQR = 1), whereas the median in tailored group was 5.33 (IQR = 0.67). The Wilcoxon test showed that the **difference was significant** with p = 0.02744.

The correlation between several metrics was also analysed, specifically to investigate

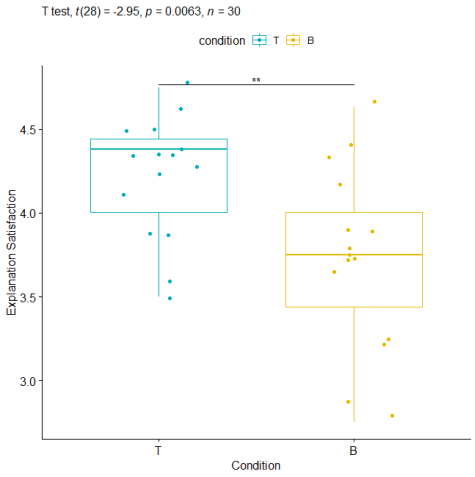

Figure 1: Mean Explanation Satisfaction Value by Agent Condition

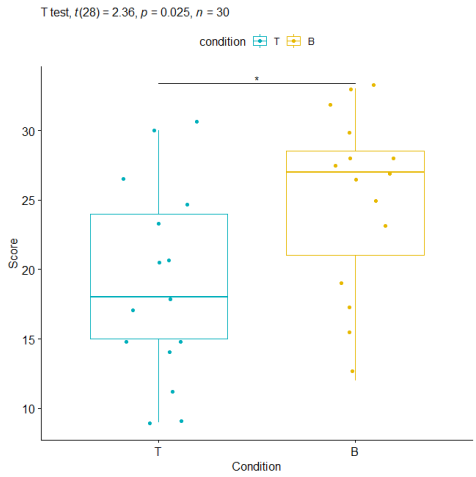

Figure 2: Mean Score Value by Agent Condition

the influence of gaming experience on the metrics. A Spearman's rank-order correlation (with significance at $\alpha = 0.05$) was run to determine the gaming experience's relationship to human moves, score, subjective workload, human idle time and simultaneous idle time. Only the subjective workload was found to correlate with gaming experience. There was a weak, negative correlation between the metrics, which was **statistically significant** (correlation coefficient = -0.3696012, p = 0.04441).

Also to overrule if high-scoring participants were more satisfied with the explanations, a Pearson correlation test (with significance at $\alpha = 0.05$) was run to determine the relationship between *explanation satisfaction and score*. No significant correlation was found (correlation coefficient = -0.2699415, p = 0.1491).

5 Discussion

This section initially reviews the results presented in Section 4 and interprets their meanings. Next, the limitations of the project are openly discussed. Finally, future studies and results of the research are indicated.

5.1 Interpretation of Results

The results obtained from the gaming experience analyses show that there is no significant difference between the two experimental groups. Correlation tests only resulted in one significant correlation between gaming experience and subjective workload. The correlation was not strong, but this confirmed a suspicion that a participant with a low gaming experience would find the game more overwhelming. Regarding other demographic information, the groups were well divided and had no significant difference, as said in Section 3.2.

The overall findings from the metrics show that there is no significant difference between trust metrics and most of the performance metrics. Similar means were expected across these metrics, as minimal changes were made to the overall game mechanics and the agent's

brain. The results show that tailoring only affected a small portion of the dependent variables. Significant differences between means exist in completeness and score metrics, which indicate that the baseline group performed better in the collaborative task, as observed in Figure 2. This is likely due to an information overload experienced by under-performing participants, which was caused by the performance-tailored explanations. Long messages caused participants to waste time reading them, while the agent was sitting idle. Under-performing participants received many tips and reminders, causing them to communicate less with the agent. This conclusion is also supported by the significant difference found in the agent moves metric. The tailored agent made fewer moves in general and contributed less because the human took a long time to respond to its messages.

The attention given to the tutorial game vastly affected the game performance of the participants. After the tutorial, each participant had a different level of understanding and in some cases, people expressed the need to play one more time before the actual game to fully understand the important game mechanics. Since this was not allowed, in future, a longer tutorial game might fill the knowledge gap between participants better. This also supports the data gathered from the subjective measures, recorded after the experiments, where the game mechanics were fully understood.

Subjective metrics reveal new findings. The collaboration fluency metric on its own did not show any significant difference, however, a deeper analysis of its sub-metrics revealed a significant difference in improvement. This could be a result of the agent explanations becoming more stable after some time in the game, especially after the first few points. The implementation had a minor downside which caused the first 1-2 minutes of the game to be seen as a poor performance since no points are gained yet. While the agent is learning about its new teammate, it is not easy to create a performance model right away, which could cause instability in explanations. Moreover, a major difference can be seen in Figure 1 between the two groups' results in explanation satisfaction. Interestingly, although objectively the participants performed poorly on the task, they were satisfied with the agent's process and explanations. We could conclude that the explanations were satisfactory and were appreciated, however, due to the time pressure, they created a separate challenge. This can be attributed to the gamified nature of the environment, where time constraints play a large role. These results can be considered unsatisfactory for USAR-HAT performance in the end but can serve as an effective solution for other collaborative task environments.

5.2 Limitations

This research was completed as a dissertation for the Bachelor of Computer Science and Engineering at TU Delft. While the project had its advantages of working together in a team under a well-established research group, it also had challenging time constraints, that limited the research output.

The two-month time constraint of the project limited the research depth of the paper at several moments in the process. The large size and variety of the field made it clear that more time is required to study all related research, answering the research question more concretely by creating a better environment. In addition to the references cited at the end, there are many more valuable studies about XAI, some of which are still being published at the time of this paper. The time constraints especially limited the main experimentation. While 30 participants is an adequate amount for a statistically correct conclusion to be made, the external factors could have been more carefully avoided by using more diverse participants. Especially the gaming experience factor of the participants could be kept under

a strict inclusion factor, which would allow the differences in the results to be only based on the tailoring factor.

Moreover, the topics of XAI and tailoring explanations are vague terms by their nature. There are many approaches for implementing user-aware solutions in HAT. Tailoring human performance was done in this research, in the way it was described in Section 3, however, this does not necessarily define the only way of achieving a performance-aware agent. This indeed also becomes a limitation, where this research on its own, cannot eliminate any other conclusions made through this specific tailoring. However, the reproducibility of the experiment, the quality of the gathered data and the depth of the tailoring, also prove significant contributions to the topic of tailoring explanations in XAI.

5.3 Future Work

The next step for the XAI community is to conduct further research on user-aware agents, both on the topic of this paper and also several other possible tailoring factors. After more results prove which factors contribute most to the overall performance and satisfaction of the user, the combination of those factors should also be researched on metrics similar to the ones used in this paper. These metrics provide an extensive understanding of the effects of the factors, which eventually will take human-agent teaming to a whole new level. While agents in these use cases become better and better at explaining themselves to their human partners, the potential for their collaboration increases.

6 Responsible Research

This research was mainly about the interaction between a human and an AI agent. The main part of the project involved an experiment with human participants. Thus, the ethical concern of the project becomes the reproducibility of the experiment and the procedure of the experiment for participants.

As stated in the Netherlands Code of Conduct for Research Integrity [35] the five principles forming the basis of integrity in research are honesty, scrupulousness, transparency, independence and responsibility. All five principles were applied in this research process and all precautions were taken to ensure that no unintentional mistakes were made.

This research has applied and received the approval of the TU Delft's Human Research Ethics Committee (HREC)[8], confirming the suitability of the research in terms of ethical implications. Before the experimentation, participants were given an 'Informed Consent Form', explaining, in detail, the purpose of the research, how to opt-out of the research and how their collected data will be stored. This allowed the participants to understand the research and their rights during or after the experimentation. In addition, the experiments were carried out in person at the TU Delft campus, in accordance with the coronavirus measures of the Dutch government at the time.

The experiment leader valued giving all participants the same experimental conditions. They were given a chance to play one round of the tutorial game and one round of the actual game, with a cheat sheet beside the experiment laptop. If the first run of the game failed after halftime, the experiment stopped and that participant's data was not used, because the participant was considered 'experienced' with the game environment (which did not happen during the experiments). Previous experience with the MATRX environment

[8] tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics

was not considered an important confounding factor, since the task of this experiment had a completely different goal and environment. At no time during the experiment, do the participants know which experimental group they are in, or what are the differences between the groups.

Statistical analysis and checks were done for all evaluation data that was gathered in the experimentation process. All data that was gathered to conclude the experiment was presented in the corresponding sections of the paper. Any data that was deemed not useful or excluded from the experiment was stated in the paper clearly. In this case, no data were excluded from the experiment.

Limitations of the experiment were clearly explained in Section 5.2, which justifies the conclusions made in this study and points out clearly what parts could be further researched.

Over the detailed explanations made in Section 3 to ensure reproducibility of the experiment, the source code and agent implementations were made public and free to use for any further research[9]. The main framework, MATRX, that was used in the implementation was also available publicly as an open-source tool. Any other tool, documentation or implementation used in this research can be made available upon request.

Aside from all the aforementioned concerns, there was nothing else identified as ethically concerning. The contact details of the experiment leader, author of this paper and supervisor of the research are also easily accessible to the participants and to any reader of this paper, to resolve future conflicts.

7 Conclusion

This study aimed to find the impact of tailoring agent explanations according to human performance on human-agent teamwork, through a game of urban search-and-rescue using the MATRX software environment. The main findings of this experiment were that this type of tailoring resulted in lower collaborative performance, but significantly higher explanation satisfaction. This indicates that the tailoring factor is valuable and must be further studied and refined to become an asset in interdependent task communication. All research questions were answered within limits. The findings of the research contribute to the larger study area of XAI and the effort to make artificially intelligent agents more understandable to their human partners. Further research on the topic could discover the best factors to use for tailoring explanations, which in the end could result in a similar conceptual framework used to explain human behaviour. This result could pave the way to a future where agents and humans work collaboratively on several tasks, with high trust in each other, satisfaction from the collaboration and increased team performance. Artificially intelligent agents can contribute a lot to a team, especially with their ability to process large amounts of data and find out the best next action, their contribution becomes more apparent and effective if they can explain the reasoning behind their conclusions. XAI has a long way to go to emulate real human explainability models, however, it is close to a breakthrough in HAT systems.

[9] github.com/canparlar/TUD-Research-Project-2022-Human-Performance

A MATRX Game Environment

Below, Figure 3 is an image of the MATRX game environment used in the controlled experiments. This is an image of the 'God' view where all obstacles and victims are visible, disregarding the view range of the human. On the left, the game map is displayed and on the right is the chat feature, which includes the first message of the agent.

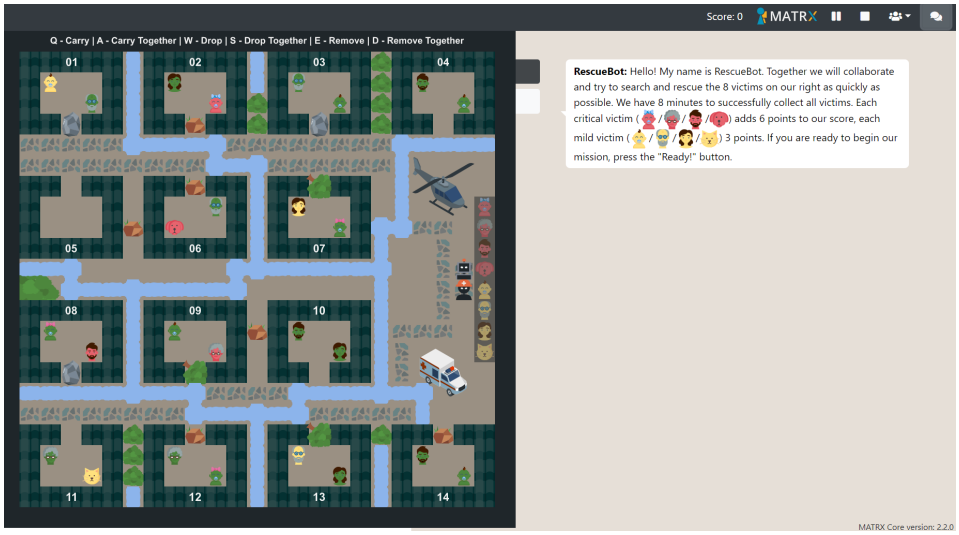

Figure 3: Screenshot of the MATRX Game Environment

References

[1] National Academies of Sciences Engineering and Medicine, *Human-AI Teaming: State-of-the-Art and Research Needs*. Washington, DC: The National Academies Press, 2022. [Online]. Available: https://doi.org/10.17226/26355

[2] G. Rampersad, "Robot will take your job: Innovation for an era of artificial intelligence," *Journal of Business Research*, vol. 116, pp. 68–74, 8 2020. [Online]. Available: https://doi.org/10.1016/j.jbusres.2020.05.019

[3] M. Mirbabaie, F. Brünker, N. R. J. Möllmann Frick, and S. Stieglitz, "The rise of artificial intelligence - understanding the AI identity threat at the workplace," *Electronic Markets*, 2021. [Online]. Available: https://doi.org/10.1007/s12525-021-00496-x

[4] J. Bradshaw, R. Hoffman, M. Johnson, and D. Woods, "The Seven Deadly Myths of "Autonomous Systems"," *Intelligent Systems, IEEE*, vol. 28, pp. 54–61, 5 2013. [Online]. Available: https://doi.org/10.1109/MIS.2013.70

[5] J. van Diggelen, J. S. Barnhoorn, M. M. M. Peeters, W. van Staal, M. L. Stolk, B. van der Vecht, J. van der Waa, and J. M. Schraagen, "Pluggable Social Artificial Intelligence for Enabling Human-Agent Teaming," *ArXiv*, 9 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1909.04492

[6] F. Santoni de Sio and J. van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account," *Frontiers in Robotics and AI*, vol. 5, 2018. [Online]. Available: https://doi.org/10.3389/frobt.2018.00015

[7] L. C. Siebert, M. L. Lupetti, E. Aizenberg, N. Beckers, A. Zgonnikov, H. Veluwenkamp, D. A. Abbink, E. Giaccardi, G.-J. Houben, C. M. Jonker, J. van den Hoven, D. Forster, and R. L. Lagendijk, "Meaningful human control over AI systems: beyond talking the talk," *CoRR*, vol. abs/2112.01298, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2112.01298

[8] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. van Riemsdijk, and M. Sierhuis, "Coactive Design: Designing Support for Interdependence in Joint Activity," *Journal of Human-Robot Interaction*, vol. 3, no. 1, pp. 43–69, 2 2014. [Online]. Available: https://doi.org/10.5898/JHRI.3.1.Johnson

[9] E. M. van Zoelen, K. van den Bosch, and M. Neerincx, "Becoming Team Members: Identifying Interaction Patterns of Mutual Adaptation for Human-Robot Co-Learning," *Frontiers in Robotics and AI*, vol. 8, 2021. [Online]. Available: https://doi.org/10.3389/frobt.2021.692811

[10] M. Johnson and A. Vera, "No AI Is an Island: The Case for Teaming Intelligence," *AI Magazine*, vol. 40, no. 1, pp. 16–28, 3 2019. [Online]. Available: https://doi.org/10.1609/aimag.v40i1.2842

[11] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019. [Online]. Available: https://doi.org/10.1016/j.artint.2018.07.007

[12] G. David, S. Mark, C. Jaesik, M. Timothy, S. Simone, and Y. Guang-Zhong, "XAI-Explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, p. eaay7120, 12 2019. [Online]. Available: https://doi.org/10.1126/scirobotics.aay7120

[13] M. Rosen, S. Fiore, E. Salas, M. Letsky, and N. Warner, "Tightly Coupling Cognition: Understanding How Communication and Awareness Drive Coordination in Teams," *The International C2 Journal*, vol. 2, no. 1, pp. 1–30, 5 2008. [Online]. Available: https://www.researchgate.net/publication/235132556_Tightly_Coupling_Cognition_Understanding_How_Communication_and_Awareness_Drive_Coordination_in_Teams

[14] N. Hanna, R. Deborah, and M. Hitchens, "Evaluating the Impact of the Human-Agent Teamwork Communication Model (HAT-CoM) on the Development of a Shared Mental Model," in *PRIMA 2013: Principles and Practice of Multi-Agent Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 453–460. [Online]. Available: https://doi.org/10.1007/978-3-642-44927-7_34

[15] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 6 2019. [Online]. Available: https://doi.org/10.1609/aimag.v40i2.2850

[16] S. T. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable Agents and Robots: Results from a Systematic Literature Review," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, 2019. [Online]. Available: https://dl.acm.org/doi/10.5555/3306127.3331806

[17] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman, "A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable," *Explainable and Transparent AI and Multi-Agent Systems*, pp. 119–138, 2021. [Online]. Available: https://doi.org/10.1007/978-3-030-82017-6_8

[18] E. Salas, J. A. Cannon-Bowers, and J. H. Johnston, "How can you turn a team of experts into an expert team?: Emerging training strategies." in *Naturalistic decision making.*, ser. Expertise: Research and applications. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, 1997, pp. 359–370. [Online]. Available: https://doi.org/10.1518/001872008X288385

[19] K. Stowers, L. L. Brady, C. MacLellan, R. Wohleber, and E. Salas, "Improving Teamwork Competencies in Human-Machine Teams: Perspectives From Team Science," *Frontiers in Psychology*, vol. 12, 2021. [Online]. Available: https://doi.org/10.3389/fpsyg.2021.590290

[20] T. A. Schoonderwoerd, E. M. v. Zoelen, K. v. d. Bosch, and M. A. Neerincx, "Design patterns for human-AI co-learning: A wizard-of-Oz evaluation in an urban-search-and-rescue task," *International Journal of Human-Computer Studies*, vol. 164, p. 102831, 8 2022. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2022.102831

[21] M. Harbers, J. M. Bradshaw, M. Johnson, P. Feltovich, K. van den Bosch, and J.-J. Meyer, "Explanation and Coordination in Human-Agent Teams: A Study in the BW4T Testbed," in *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 3, 2011, pp. 17–20. [Online]. Available: https://doi.org/10.1109/WI-IAT.2011.83

[22] M. Harbers, J. M. Bradshaw, M. Johnson, P. J. Feltovich, K. van den Bosch, and J.-J. Meyer, "Explanation in Human-Agent Teamwork," *Coordination, Organizations, Institutions, and Norms in Agent System VII*, pp. 21–37, 2012. [Online]. Available: https://doi.org/10.1007/978-3-642-35545-5_2

[23] E. Salas, D. E. Sims, and C. S. Burke, "Is there a "Big Five" in Teamwork?" *Small Group Research*, vol. 36, no. 5, pp. 555–599, 10 2005. [Online]. Available: https://doi.org/10.1177/1046496405277134

[24] M. P. D. Schadd, T. A. J. Schoonderwoerd, K. van den Bosch, O. H. Visker, T. Haije, and K. H. J. Veltman, ""I'm Afraid I Can't Do That, Dave"; Getting to Know Your Buddies in a Human-Agent Team," *Systems*, vol. 10, no. 1, 2022. [Online]. Available: https://doi.org/10.3390/systems10010015

[25] K. R. McKee, X. Bai, and S. T. Fiske, "Warmth and competence in human-agent cooperation," in *21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)*, 1 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2201.13448

[26] G. Klein, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich, "Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity," *IEEE Intelligent Systems*, vol. 19, no. 6, pp. 91–95, 11 2004. [Online]. Available: https://doi.org/10.1109/MIS.2004.74

[27] M. Johnson, C. Jonker, B. van Riemsdijk, P. J. Feltovich, and Bradshaw Jeffrey M, "Joint Activity Testbed: Blocks World for Teams (BW4T)," in *Engineering Societies in the Agents World X*, H. Aldewereld, V. Dignum, and G. Picard, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 254–256.

[28] S. J. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Prentice Hall, 2002.

[29] G. J. Lematta, P. B. Coleman, S. A. Bhatti, E. K. Chiou, N. J. McNeese, M. Demir, and N. J. Cooke, "Developing Human-Robot Team Interdependence in a Synthetic Task Environment," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 1503–1507, 2019. [Online]. Available: https://doi.org/10.1177/1071181319631433

[30] J. van der Waa, S. Verdult, K. van den Bosch, J. van Diggelen, T. Haije, B. van der Stigchel, and I. Cocu, "Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations," *Frontiers in Robotics and AI*, vol. 8, 2021. [Online]. Available: https://doi.org/10.3389/frobt.2021.640647

[31] J. van Diggelen, M. Neerincx, M. Peeters, and J. M. Schraagen, "Developing Effective and Resilient Human-Agent Teamwork Using Team Design Patterns," *IEEE Intelligent Systems*, vol. 34, no. 2, pp. 15–24, 2019. [Online]. Available: https://doi.org/10.1109/MIS.2018.2886671

[32] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for Explainable AI: Challenges and Prospects," *XAI Metrics*, 12 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1812.04608

[33] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Advances in Psychology*, vol. 52, no. C, pp. 139–183, 1 1988. [Online]. Available: https://doi.org/10.1016/S0166-4115(08)62386-9

[34] G. Hoffman, "Evaluating Fluency in Human-Robot Collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 209–218, 2019. [Online]. Available: https://doi.org/10.1109/THMS.2019.2904558

[35] KNAW, NFU, NWO, TO2-Federatie, Vereniging Hogescholen, and VSNU, "Netherlands Code of Conduct for Research Integrity," DANS, Tech. Rep., 2018. [Online]. Available: https://doi.org/10.17026/dans-2cj-nvwu