Agent Failure and Trust Repair in Human-Agent Teams

Interdependence Impact on Trust Repair Strategy and Collaboration Fluency in Human-AI Team

Tauras Narbutas

Supervisors: Myrthe Tielman, Ruben Verhagen

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Tauras Narbutas
Final project course: CSE3000 Research Project
Thesis committee: Myrthe Tielman, Ruben Verhagen, Ujuwal Gadiraju

Abstract

Interdependence relationships between humans and agents play a crucial role in the collaborative AI field. This research paper examines the impact of interdependence on trust violation, trust repair strategies, and collaboration fluency in human-AI teams. It compares independent cooperation and required interdependence approaches, focusing on collaborative AI, trust dynamics, and collaboration fluency. The paper presents a user study involving 30 participants in different interdependence conditions, analyzing data from trust surveys, collaboration fluency surveys, performance metrics, and AI agent idle time. The findings enhance understanding of human-agent collaboration dynamics and inform the design and implementation of collaborative AI systems.

1 Introduction

The interest in the dynamics of human-agent teams has been growing in recent years as some of the studies have shown the potential of human and AI systems collaboration in fields such as surgeries or firefighting [1; 2; 3]. In such circumstances, human-AI teams rely on interdependence, which entails mutual reliance and influence, where both parties depend on each other's contributions and interact collaboratively to accomplish shared objectives or tasks [1; 4; 5]. Interdependence can arise from a lack of capacity of either humans or agents (hard constraints), or simply being more efficient when working together (soft constraints) [1; 6]. However, trust violations can occur in case one of the parties fails to accomplish the assigned task or provides a faulty input [7]. It has been concluded that the combination of both expressing regret and providing an explanation is an effective trust repair strategy [8].

On the other hand, many of the tasks such as the previously mentioned firefighting example, where human interacts with an agent to save the victims from the burning building, can compose of a variety of constraints. For example, the human may choose to either direct the agent to remove certain obstacles or collaborate with the AI system to achieve this goal faster. Therefore, analyzing the impact of the interdependence relationship in human-agent teams ought to increase the likelihood of successful trust repair and facilitate collaborative efforts. *Subsequently, this research paper aims to complement the prior research on trust repair strategies and investigate how the required interdependence relationship arising from a lack of human and robot capacities affect*

1. *the trust violation*
2. *the trust repair*
3. *the collaboration fluency*

in human-agent teams compared to an independent cooperation (baseline).

This research paper is structured as follows. The Background section outlines an overview of collaborative AI, trust violation and repair strategies as well as collaboration fluency while providing some insights into related work regarding these topics. The Methodology outlines the key fragments of the user study as well as the metrics used to analyze it. Ethics and reproducibility are discussed in the Responsible Research section. The Results section presents the findings of this research regarding different measures. The Discussion section analyzes trust repair, collaboration fluency, limitations, and suggests future work. Finally, the conclusion section summarizes the key findings and contributions of the research.

2 Background

This section provides an in-depth overview of the topics including collaborative AI, trust violation and repair strategies as well as the collaboration fluency between humans and agents. Additionally, prior research is analyzed in the following subsections in order to provide sound insights and persuade the reader of the importance of interdependence impact analysis on these concepts.

2.1 Collaborative AI

While in most other AI areas the goal is usually implementing independent systems, collaborative AI is rather focused on developing autonomous agents that have the characteristics of being observable, predictable, and directable by humans [9]. Here, the AI systems are integrated into the setting of partnership meaning that along with humans they partake in a task to accomplish common objectives. Even though continuous learning facilitated through accumulating and processing huge amounts of data remains an important factor, humans' input and guidance often influence agents' performance. Subsequently, humans rely highly on AI agents to achieve tasks optimally. As a reason, in the context of collaborative AI, interdependence relationships arise as one of the key attributes. They can be classified into several categories:

- Required relationships: High interdependence between human and AI agent, crucial for achieving shared objectives.
- Opportunistic relationships: Utilizing AI capabilities without obligatory collaboration.
- Complementary relationships: Combining strengths of human and AI for synergistic benefits.
- Mixed relationships: Combination of different interdependency modes based on tasks or context.

Depending on the area, AI agents can support individuals by performing tasks such as information gathering, data analysis, customer service, and physical work. This allows people to focus on more advanced responsibilities that demand qualities like leadership, creative thinking, and judgment [3]. It is, however, important to understand that the collaborative AI field faces many challenges such as adequately modeling and interpreting mutual intentions and actions or managing the expenses associated with coordinated efforts [2]. For this reason, understanding the interdependence between humans and AI agents is crucial to comprehend the complexities of their cooperation and its implications for trust as well as collaboration fluency dynamics. This paper focuses on comparing two opposite approaches of human-AI agent teaming - independent versus required.

2.2 Trust Violation and Repair Strategies

Trust is a crucial element in human-agent teams since it affects some of the most fundamental components of collaboration performance [10]. This includes cooperation fluency, mutual coordination, as well as communication between the parties. Trust is composed of factors such as competence and willingness that human and agent perceives of one another. It is important to maintain these qualities throughout the collaboration phase to achieve shared goals in a reliable and competent manner. However, trust violations occur due to failures to realize a certain task or when performing misleading actions, therefore, decreasing trust between the individuals [7]. For this reason, restoring confidence in one another is a key factor in maintaining effective cooperation with regard to the collaborative AI field.

Prior research concluded that feedback is considered to have a significant effect on trust repair [11] [12]. They involve mechanisms that focus on the consequences of agent failures and work towards rebuilding trust between humans and AI agents. It turns out that providing an explanation that caused a misleading action and expressing regret about it are the key components leading to the most effective way of restoring trust [8]. On the other hand, as mentioned in the Collaborative AI section, interdependence relationships are the fundamentals of this area so it is also important to investigate how they might affect the adjustment of confidence between the parties. This research will in fact address this question in order to supplement the antecedent investigations.

2.3 Collaboration fluency

Collaboration fluency investigates how smoothly and efficiently human-agent teams interact and work together toward achieving common tasks. The metrics for this examination include capabilities, effective communication, mutual understanding, and synchronized decision-making processes between the parties. Understanding the specific factors that contribute to these components, therefore, affecting the overall collaboration fluency is an important yet hard task. In fact, it may highly depend on the specific environment in which humans and agents interact with one another as well as the common goal they are aiming to achieve.

Previous research has made significant progress in pointing out some of the key factors which might affect collaboration fluency [13]. It turns out that the idle time of either human or agent, which represents the percentage of time that an individual is not active (waiting for a response), is a good objective measure investigating communication efficiency. The same research also claims that concurrent activities and functional delays can also objectively estimate collaboration fluency. However, similarly as discussed in the Trust subsection, further investigation should be conducted to address the importance of interdependence on the effectiveness of human-agent cooperation by applying both the objective metrics as well as the subjective measures of user study questionnaires.

3 Methodology

This section presents the methodology used to investigate the interdependence impact on trust dynamics and collaboration fluency in human-AI teams. The discussion includes an overview of the user study conducted to collect the necessary data as well as the measures used for analysis.

3.1 Design

The questionnaires filled out by the participants during the user study were considered a primary method for gathering data and addressing the research question at hand. For the trust dynamics analysis, a 3 (trust survey conducted prior to violation [T1], after violation [T2], and after repair [T3]) x 2 (baseline and required interdependence conditions) mixed-design was used. After completing the game, a single questionnaire was completed by the participants to assess collaboration fluency. In addition to these subjective measures, the metrics regarding human and AI-agent actions were logged to objectively analyze the interdependence impact both on trust and cooperation efficiency.

3.2 Participants

The user study involved a total of 30 participants who were primarily recruited from TUDelft and the Uber office in Amsterdam. The number of participants was assured to be equally distributed for both of the investigated interdependence conditions (baseline and required). At the beginning of each survey, participants were required to provide their demographic information. Based on this information, condition assignments were determined to ensure a balanced representation across conditions, therefore, maintaining fairness and reducing bias in the study.

3.3 Hardware and Software

The personal laptops of the research conductors were used to carry out the user study. Each participant was introduced to two open tabs one of which contained the questionnaire and the other - the search and rescue mission UI. The questionnaires were designed using the robust survey creation tool Qualtrics, whereas the game dynamics were implemented using the human-agent teaming rapid experimentation software package MATRX.

3.4 Environment and Task

The participants interacted with a simulated scenario involving a search and rescue mission in a flooded town. The UI of the game environment is presented in Figure 1, however, when performing an actual task, humans were only able to see the objects that were no more than one block away. In total, there were 8 victims comprising 4 mildly (yellow-colored) and 4 critically (red-colored) injured individuals, each worth 6 and 3 points respectively (the max score was 26). Note that there was no need to rescue the non-injured (green-colored) persons. All of the victims were distributed among 14 different areas, some of which were blocked by either stone, tree, or rock. The communication between the human participant and the AI agent was facilitated through a chat interface, depicted in Figure 2. The AI agent utilized an algorithmic approach to adhere to the human decision-making process during the game.

The task involved collaborative efforts between the human participant and the AI agent to rescue all of the injured victims while searching areas and removing obstacles. There were two different interdependence conditions implemented for this task - baseline and required. In the case of the baseline, both the human and the agent could only perform all of the actions independently, whereas only joint actions were allowed in the required condition. It is important to understand that the time taken to pick up a victim or remove an obstacle was 5 times lower in case a human was performing it together with the AI agent. Another significant factor was heavy rain, which occurred 3 times throughout the mission at 2, 4, and 6-minute marks (the game had a total duration of 10 minutes). The human participant had to be present in one of the specified areas for at least 10 seconds. Failure to do so resulted in a loss of 10 points. Before the first and the third storm, the AI agent gave the correct advice for the human to seek shelter. However, the second piece of advice was incorrect and so the agent attempt to repair the trust by expressing regret and providing an explanation [8].

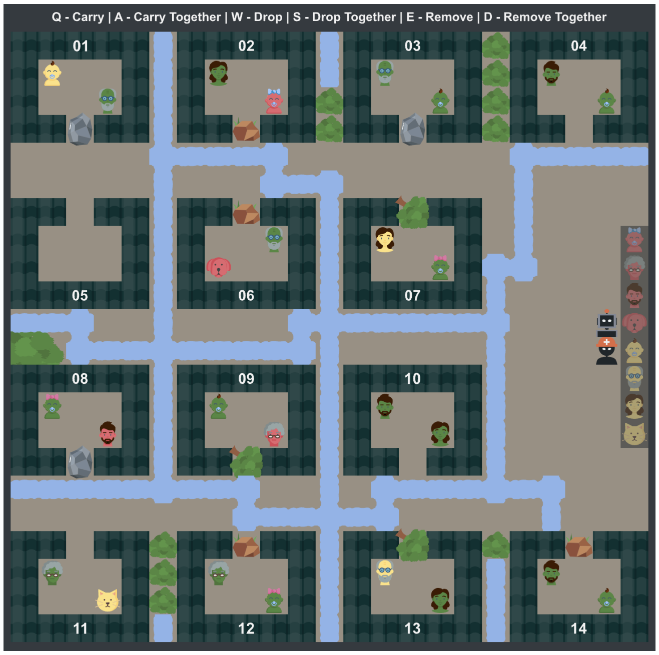
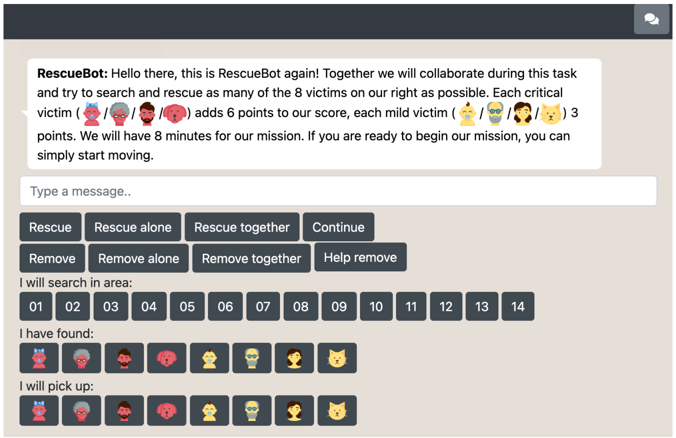

Q - Carry | A - Carry Together | W - Drop | S - Drop Together | E - Remove | D - Remove Together
01
02
03
04
05
06
07
08
09
10
11
12
13
14

Figure 1: UI of the search and rescue game



RescueBot: Hello there, this is RescueBot again! Together we will collaborate during this task and try to search and rescue as many of the 8 victims on our right as possible. Each critical victim (/ / /) adds 6 points to our score, each mild victim (/ / /) 3 points. We will have 8 minutes for our mission. If you are ready to begin our mission, you can simply start moving.

Figure 2: Chat interface used for the human-agent communication

3.5 Measures

In this section, the measures used to analyze the results of the study are presented. They include trust and collaboration fluency surveys as well as objective metrics such as performance, AI agent idle time, number of messages sent, and human location during the storm. Each of the subsections indicates how the measure has been collected and what it was used for.

3.5.1 Trust

To assess the level of human trust in the AI agent, a questionnaire was administered to participants. This questionnaire is equivalent to the one used in the research for investigating the effectiveness of different trust repair strategies [8]. The questionnaire included items related to trust, such as predictability and reliability as well as other subjective metrics. For the trust questionnaire, 5-point Likert scales ranging from 'I disagree strongly' to 'I agree strongly' were used. The responses were converted into comparable numeric values between 1 and 5 for further analysis.

3.5.2 Collaboration Fluency

To measure collaboration fluency in the human-AI team, a questionnaire, also used in another study for evaluating the human-robot team fluency [13], was utilized. The questionnaire included items that assessed the ease of communication, coordination, and cooperation between the human and the AI agent. For the collaboration fluency questionnaire, 7-point Likert scales ranging from 'I disagree strongly' to 'I agree strongly' were used. The responses were then converted into comparable numeric values between 1 and 7 so that these subjective metrics could then be easily used for further analysis.

3.5.3 Performance

Performance is one of the objective metrics that was evaluated based on two main factors: score and task completeness. The score represented the number of injured victims successfully rescued by the human-AI team. Task completeness measured the extent to which the team accomplished the overall objective of the rescue mission. While completeness could be 100%, the score might suffer from a human being exposed

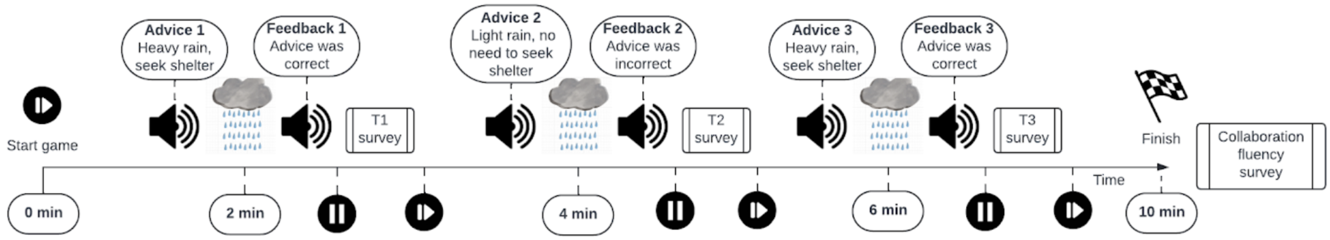


Figure 3: Schematic timeline depicting the experiment. Participants were allotted a total of 10 minutes to complete the game. After each storm, the game was paused (at minutes 2, 4, and 6) and participants had to fill out trust surveys (T1, T2, and T3). Upon game completion, participants also completed the collaboration fluency survey

to rain as explained in the Task section above. Additionally, the time taken to complete the task was recorded as a performance measure.

3.5.4 Agent Idle Time

The idle time of the AI agent was logged as another objective measure throughout the experiment. This metric captured the duration while AI agent remained inactive, meaning that it was not contributing to the task. It was checked for in case valuable insights were provided into the human's level of engagement to utilize the agent within the collaborative setting.

3.5.5 Number Human-Sent Messages

The total number of messages exchanged between the human and the agent is another objective measure. It was logged to gain a better understanding of communication fluency within the human-AI team. More specifically, human engagement in collaboration with the agent was checked in this scenario.

3.5.6 Human Location During Storm

This is another objective metric that captures the human's location within the environment during periods of heavy rain. It was logged to assess the compliance of the human with the task requirements and the impact of adverse weather conditions on the team's performance. Moreover, this metric can also be indicative of communication efficiency, as humans may tend to check the chat more frequently depending on the condition.

3.6 Procedure

Before starting the task, participants were required to read and complete the informed consent form in the Qualtrics survey, followed by providing their demographic information. Afterward, the participants underwent a tutorial where they were familiarized with the chat interface, learned to navigate the map and make use of game the controls as well as gained an understanding of how to interact with the AI agent on their team. Once the introduction of the game and the survey has been completed, the participants received instructions for the official task of collaborating with the AI agent in a search and rescue mission. The experimenter also addressed participants' questions and provided them with a cheat sheet specific to their assigned condition, while ensuring that the research questions remained undisclosed. Throughout the task, the game was periodically paused after each storm for participants to answer Qualtrics survey questions regarding trust in the AI agent they collaborated with. Once ready, the experiment commenced with the experimenter refreshing the human view interface. After the game finished, the participants had to fill out the last survey regarding collaboration fluency. Additionally, Figure 3 showcases the procedure of the user study by depicting the schematic timeline of the experiment.

4 Responsible Research

In scientific papers as such, it is important to consider the ethical implications as well as the reliability of the research. This section provides an overview of the ethical considerations and reproducibility characteristics of the study conducted on the interdependence impact on trust violation and repair strategy, and collaboration fluency in human-AI teams. The subsections below outline the measures taken to confirm responsible research practices.

4.1 Ethics

Ethical considerations are crucial in research that involves user studies with human participants. For this reason, the informed consent form was presented to all participants before proceeding with the survey alongside providing them with procedures, and potential risks. Participants were also informed about their voluntary participation and the right to withdraw from the study at any time. Clear instructions were provided, and participants were given the opportunity to ask questions and seek clarification. On the other hand, both the purpose and the hypothesis of this research have been kept a secret in order to maintain participants' integrity.

The study also considered potential biases and ensured fairness in participant recruitment and assignment to interdependence conditions. Efforts were made to recruit participants from diverse backgrounds to ensure the generalizability of the findings. By employing a mixed design and random assignment, the study aimed to reduce bias and maintain an unbiased representation of interdependence conditions.

4.2 Reproducibility

Reproducibility promotes transparency and validation of findings. This study ensures such qualities by describing the methodology, materials, and measures used. The user study procedures have been outlined, including trust surveys, collaboration fluency questionnaires as well as specified objective metrics, facilitating replication and validation of the results. To summarize, a clear framework for reproducing the study in a similar context has been provided.

The open-source MATRX software has been used, therefore, enabling others to conduct comparable experiments. With its capabilities to simulate collaborative scenarios and facilitate human-AI interactions, this study can be leveraged for reproducibility and further research in the field. In addition to that, other researchers can also access the implementation of the project on https://github.com/mawakeb/CSE3000-2023-trust-repair and replicate or continue the study.

5 Results

In this section, the statistical analysis of the data collected during the user study is presented. In particular, the significance of different interdependence conditions (baseline and required) was investigated on each of the measures described in Section 3.6. The statistical programming language R was used to conduct the analysis and visualize data distributions.

5.1 Trust

The robust two-way mixed ANOVA test was conducted to investigate the interdependence impact on human trust in the AI agent over time. This test included the between-subject factor (interdependence) condition (baseline versus required) and the within-subject factor time (trust after correct advice T1 versus after violation T2 versus after repair T3). The robust version of the test was chosen to account for the validity of the results as there were some of the following data normality violations:

- The significant outliers were checked for by visualizing the data using a box plot and the *identify_outliers()* [rstatix package] methods. For the required condition, three outliers were identified, two of which were extreme.
- The normality of the variables for each combination of factor levels was checked using the Shapiro-Wilk normality test and inspecting the QQ plots. These tests concluded that for the required condition, trust values at time T3 are not normally distributed as $p = 0.0144 < 0.05$.
- Levene's test was used to check for the homogeneity of variance of the between-subject factor condition, which resulted in violation at T3 as $p = 0.00738 < 0.05$.

Having this considered, the resulting values of the robust two-way mixed ANOVA test are presented in Table 1. It can be seen that there was a statistically significant interaction effect between condition and time factors on the trust as $p = 0.005 < 0.05$. However, the condition effect was insignificant as $p = 0.477 > 0.05$. For this reason, only the simple main effects of the time variable were investigated by computing the pairwise comparisons between the time points at each condition. The outcomes concluded that the effect of time was significant on the required condition in all cases, whereas for the baseline condition, all $p-values$ were larger than 0.05. The test results for the required condition are presented below:

Effect	value	p
cond	0.5184	0.477
time	56.6557	0.001
cond:time	13.3104	0.005

Table 1: Values resulting from the Two-way Mixed ANOVA test for the dependent variable trust

- T1 vs T2: $p = 0.00000405 < 0.5$.
- T1 vs T3: $p = 0.000855 < 0.5$.
- T2 vs T3: $p = 0.0000435 < 0.5$.

Additionally, the relationship between trust and time per condition has been illustrated by plotting the estimated marginal means in Figure 4.

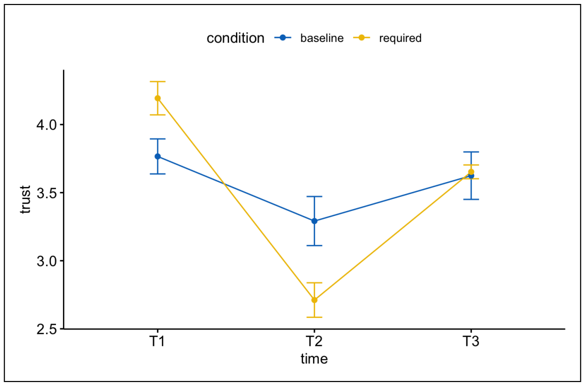


Figure 4: Estimated marginal means illustrating the relationship of trust (Y-axis) and time (T1, T2, and T3) (X-axis) per condition

5.2 Collaboration Fluency

The Kruskal-Wallis test, which is a non-parametric version of the one-way mixed ANOVA, was used to evaluate the interdependence significance with regard to collaboration fluency. A pipe-friendly *kruskal_test()* function [rstatix package] has been applied to analyze the resulting scores of the user surveys. It yielded that there were no significant differences between the average fluencies in the two (required and baseline) experimental interdependence conditions as $p = 0.803 > 0.05$. The box plot has been depicted in Figure 5, which showcases the collaboration fluency score in terms of minimum, maximum, and median as well as first and third quartiles per condition.

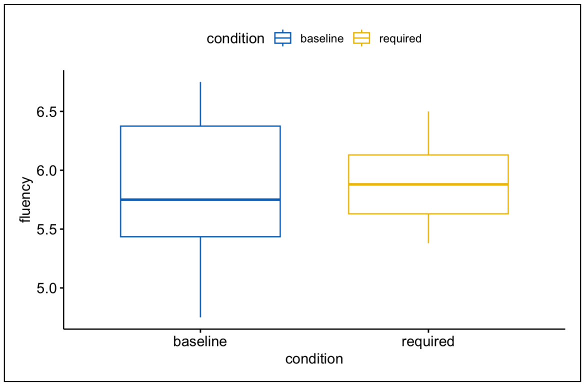


Figure 5: Box plot depicting the collaboration fluency (Y-axis) per condition (X-axis)

5.3 Performance

As stated in the Methodology section, performance is measured in terms of three metrics: completeness, score, and time taken to finish the task. Below, the box plot has been depicted for each of the metrics per condition in Figures 6, 7, and 8. Similarly to investigating the effects of interdependence conditions on collaboration fluency, each of the 3 performance measures was also analyzed using the Kruskal-Wallis test. The test concluded that both completeness and time taken to finish the task were significantly affected by the interdependence condition, having p-values equal to 0.0198 and 0.00123 respectively (< 0.05 in both cases). However, the final score seems to be unaffected as the statistics yielded that $p = 0.298 > 0.05$. Next, the Kruskal-Wallis test effect size was inspected for the two impacted measures (completeness and time) by computing the eta squared, based on the H-statistic. It is calculated as $eta2[H] = (H - k + 1)/(n - k)$, where H is the test statistic, k is the number of conditions, and n is the total number of observations [14]. In both cases, a large effect was discovered on the differences between conditions - $eta2[H] = 0.158$ for completeness and $eta2[H] = 0.338$ for the time taken to finish the task.

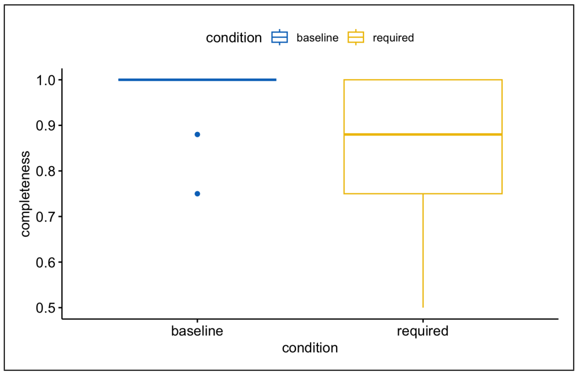


Figure 6: Box plot depicting the completeness of the game in proportion between 0 and 1 (Y-axis) per condition (X-axis)

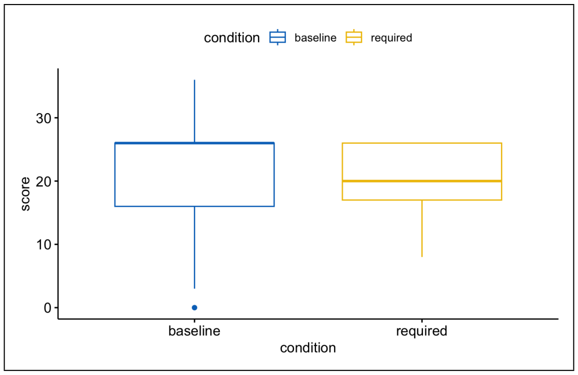


Figure 7: Box plot depicting the final score after finishing the game (Y-axis) per condition (X-axis)

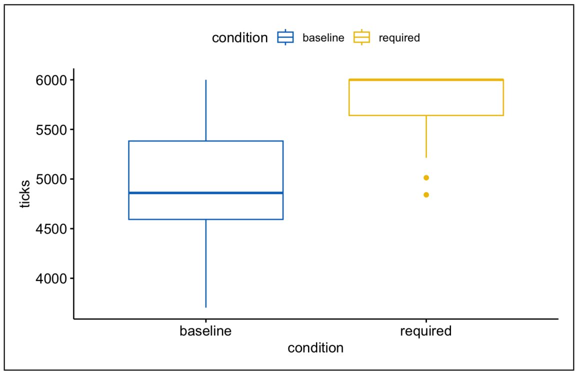


Figure 8: Box plot depicting statistics of the time that was needed to finish the game (Y-axis) per condition (X-axis)
. Here, the time is depicted in the number of ticks where 10 ticks are equivalent to 1 second

5.4 Agent Idle Time

Agent idle time is another measure that was analyzed using the non-parametric Kruskal-Wallis test. The resulting statistics, which were computed following the steps previously presented in the Performance section, yielded that $p = 0.00000824 < 0.05$ and $eta2[H] = 0.674$. These outcomes conclude that the interdependence condition has a significant effect on the total agent idle time per game. Additionally, the box plot presented in Figure 9 supplements the Kruskal-Wallis test results as it can be seen that the two conditions have different distributions.

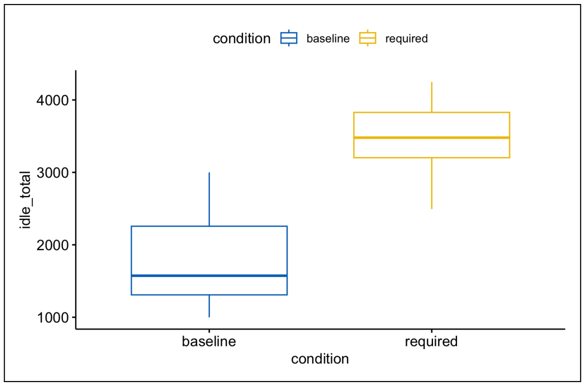


Figure 9: Box plot depicting the total idle time in the number of ticks of AI agent throughout the game (Y-axis) per condition (X-axis)

5.5 Number of Human-Sent Messages

The box plot visualized in Figure 10 presents the total number of human-sent messages per game depending on the condition. It can be seen that the two means are quite similar, yet the variance is much higher for the baseline. The Kruskal-Wallis test was performed here in order to analyze the data statistically and compute defined outcomes. It turns out that the resulting $p = 0.0963 > 0.05$, meaning that even though there are differences in the plots, the condition impact on the number of human-sent messages should be considered insignificant.

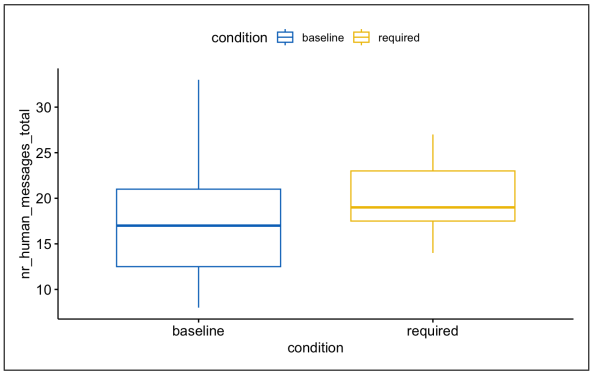


Figure 10: Box plot depicting the total number of human-sent messages throughout the game (Y-axis) per condition (X-axis)

5.6 Human Location During Storm

To compare the data gathered by logging human location during storms per condition, the percentage of how often the participants followed the agent-given suggestions was computed. To be more specific, at times T1 and T3 humans must have been in one of the shelters in case they followed the advice. However, the data gathered at T2 is disregarded in the calculations as humans could have chosen to be anywhere as they were suggested to ignore the upcoming rain. Having that said, for the baseline condition humans were hiding in shelter during storms at times T1 and T3 only $66,67\%$ of the time while for the required - $96,67\%$. These results conclude that interdependence conditions had a significant influence on human location during storms.

6 Discussion

In this section, the results regarding the impact of interdependence relationships on trust violation, trust repair, and collaboration fluency in human-agent teams are discussed. To be more specific, each of the statistical outcomes is compared between baseline and required conditions. Furthermore, the limitations of the user study as well as the proposed future work are presented.

6.1 Trust Violation and Repair

First of all, it is important to note that the trust repair strategy of giving explanations and expressing regret for agent failures was an important factor in rebuilding trust and so these findings are consistent with previously conducted research [8]. Additionally, the results of this user study provide valuable insights into the interdependence relationship impact on trust violation and trust repair in human-agent teams. In the required experimental condition, participants were highly reliant on the agent to complete all of the tasks in a joint manner. When their own capacities were insufficient or mismatched with the capabilities of the agent, participants encountered difficulties in successfully accomplishing their objectives. These challenges resulted in a perceived violation of trust as concluded by the statistically robust two-way ANOVA test, presented in Section 5.1. On the contrary, there was no breakdown in human trust with regard to the baseline condition when the agent gave faulty advice. It turns out that hard constraints had a significant effect on participants questioning the reliability and competence of the agent compared to when there were no constraints.

Furthermore, trust repair became crucial to restore the human perception of confidence and trustworthiness of the agent. The conducted user study yielded that trust repair was more challenging in the baseline condition compared to the required condition. Participants required more evidence of the agent's competence and reliability before they were willing to trust it again. The process of trust repair involved increased monitoring and verification of the robot's actions, seeking additional confirmations, and requesting explanations for its decisions. Trust repair in the required interdependence condition took more time and effort compared to the independent cooperation condition, indicating the impact of interdependence on rebuilding trust in human-agent teams.

6.2 Collaboration Fluency

The interdependence condition on the collaboration fluency in human-agent teams was investigated by analyzing the corresponding questionnaire results. It turns out that even though the mean is higher for the required interdependence relationship, the condition is considered to have no significant influence on the cooperation effectiveness as concluded by the Kruskal-Wallis test, performed in Section 5.2. However, important to understand that the limitations of the agent's capacities may have affected the overall performance flow, as participants had to invest additional effort in clarifying instructions, adjusting strategies, and compensating for the limitations. This increased cognitive load and reduced the efficiency and effectiveness of the collaboration process. The findings suggest that a lack of human and robot capacities in required interdependence relationships can be a significant barrier to achieving optimal collaboration fluency in human-agent teams.

Furthermore, the performance of the teamwork was analyzed using the objective metrics, logged throughout the experiment. It turns out the condition had a significant influence on the completeness of the game as well as the time needed to finish the task. The impact of the score, on the other hand, was negligible. This observation is closely tied to the finding that in the baseline condition, humans were much less inclined to follow the correct suggestions from the agent to seek shelter from heavy rain compared to situations where the required interdependence relationship was emphasized, as discussed in Section 5.6. Therefore, despite collaboration being deemed more effective in the baseline, the presence of required interdependence had a noteworthy positive influence on the fluency of human-agent communication.

Finally, a statistically robust Kruskal-Wallis test was conducted to analyze the total idle time of the AI Agent. The test yielded significant results, indicating that in the independent team dynamics, the agent had significantly shorter idle periods compared to the required interdependence condition. However, note that these findings may be rather influenced by the dynamics of the game itself rather than the specific interdependence condition. This is because, for the required condition, the agent will never pick up the victim alone, and so it will idle until the response is provided.

6.3 Limitations

While our study provides valuable insights into the impact of required interdependence on trust violation, trust repair, and collaboration fluency, it is important to acknowledge certain limitations. Firstly, the experimental setup simulated specific scenarios and tasks, which may not fully capture the complexities of real-world human-agent interactions. The generalizability of our findings to different domains and contexts should be examined in future research.

Secondly, our study focused on the effects of limited human and robot capacities on trust and collaboration fluency. Other factors, such as team dynamics, individual differences, and prior experience [15], may also influence these outcomes and should be considered in future investigations. Additionally, the specific tasks and levels of complexity in our study may have influenced the results. Varying the types and levels of tasks could provide a more comprehensive understanding of the impact of required interdependence on trust and collaboration fluency.

Lastly, the sample size in our study was relatively small, which could limit the generalizability of our results. Replicating this research with larger and more diverse participant groups would enhance the reliability and validity of the findings. Future studies should aim to recruit larger and more diverse samples to ensure the robustness of the results and to capture a wider range of perspectives and experiences.

6.4 Future Work

Future research should investigate communication strategies and feedback mechanisms to restore trust after a violation, including approaches such as providing explanations and enabling bidirectional communication. Training interventions aimed at improving collaboration fluency in interdependent teams should also be explored to mitigate the negative effects of interdependence on teamwork. Understanding individual factors like trust propensity and prior technology experience can provide insights into trust violation, repair, and collaboration fluency for personalized approaches. Conducting studies in real-world contexts (e.g., healthcare, emergency response) can offer practical insights and domain-specific guidelines for designing collaborative systems.

This study highlights the importance of aligning capacities between humans and robots to foster trust and collaboration fluency. Required interdependence relationships increase the likelihood of trust violation and pose challenges for trust repair. Collaboration fluency is affected by capacity limitations, resulting in communication and coordination difficulties. Future research should address these limitations by investigating diverse contexts, larger participant samples, and communication strategies, training interventions, and individual factors. Overall, this study contributes to the knowledge of human-agent teamwork and informs the design of collaborative systems to enhance trust, collaboration, and team performance.

7 Conclusions

This research paper investigated the impact of interdependence on trust violation, repair strategies, and collaboration fluency in human-AI teams, comparing independent cooperation with required interdependence. The findings highlight trust as a crucial element influencing collaboration performance, including factors like cooperation fluency and communication. Effective strategies such as explanations and expressing regret were found to repair trust.

The concept of interdependence is essential in collaborative AI, with various types of relationships identified. Understanding the interdependence between humans and AI agents is crucial for comprehending cooperation complexities and their implications for trust and collaboration fluency. Collaboration fluency refers to smooth interaction, influenced by metrics such as communication efficiency and concurrent activities. The research employed a user study with participants in different conditions, using a simulated search and rescue mission scenario and trust and collaboration fluency surveys, along with objective performance measures.

Overall, this research contributes to the understanding of interdependence in human-AI teams, shedding light on the dynamics and challenges of collaboration. The findings have implications for developing more effective human-AI systems across domains, including high-risk areas like firefighting and surgeries. Further research can expand upon these findings to explore additional factors and contexts that influence interdependence and its impact on collaborative AI.

References

[1] Johnson M., & Vera, A. (2019). No AI is an island: the case for teaming intelligence. AI magazine, 40(1), 16-28.

[2] Klien, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a" team player" in joint human-agent activity. IEEE Intelligent Systems, 19(6), 91-95.

[3] H. James Wilson, Paul R. Daugherty (2018). Collaborative Intelligence: Humans and AI Are Joining Forces. Harvard Business Review, 114–123.

[4] Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. Journal of Human-Robot Interaction, 3(1), 43-69.

[5] Verhagen, R. S., Neerincx, M. A., & Tielman, M. L. (2022). The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. Frontiers in Robotics and AI, 9, 243.

[6] Johnson M, Vignati M, Duran D. (2018). Understanding human-autonomy teaming through interdependence analysis. Symposium on human autonomy teaming

[7] Vries, P. W. D., Midden, C., Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. International Journal of Human-Computer Studies, 58(6), 719–735.

[8] Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. Autonomous agents and multi-agent systems, 35(2), 1-20.

[9] Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, M. B., and Sierhuis, M. (2014). Coactive Design: Designing Support for Interdependence in Joint Activity. Human-Robot Interaction, 3(1), 43-69.

[10] A. Freedy, E. DeVisser, G. Weltman and N. Coeyman (2007). Measurement of trust in human-robot collaboration. International Symposium on Collaborative Technologies and Systems, 106-114.

[11] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., Beck, H. P. (2003). The role of trust in automation reliance. International Journal of Human-Computer Studies, 58(6), 697-718.

[12] Vries, P. W. D., Van Den Berg, S. M., Midden, C. (2015). Assessing technology in the absence of proof: Trust based on the interplay of others opinions and the interaction process. Human Factors, 57(8), 1378–1402.

[13] Hoffman, G. (2019). Evaluating fluency in human-robot collaboration. IEEE Transactions on Human-Machine Systems, 49(3), 209-218.

[14] Tomczak, Maciej T., and Ewa Tomczak. (2014). The Need to Report Effect Size Estimates Revisited. an Overview of Some Recommended Measures of Effect Size. Trends in SportSciences.

[15] S. Tolmeijer, U. Gadiraju, R. Ghantasala. A. Gupta, A. Bernstein (2021). Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. UMAP '21, 77–87.