The Influence of Interdependence on Trust Repair in Human-Agent Teams

Comparing the Effectiveness of Trust Repair Strategies in Full Independence and Complementary Independence

Cherin Kim[1]

Supervisor(s): Myrthe Tielman[1]**, Ruben Verhagen**[1]

[1]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: Cherin Kim
Final project course: CSE3000 Research Project
Thesis committee: Myrthe Tielman, Ruben Verhagen, Ujwal Gadiraju

Abstract

As autonomous systems are increasingly integrated as a team member for collaborative tasks, trust in human-agent teams (HAT) becomes crucial to foster success. In many real world scenarios, trust violations are expected, thus demanding the use of trust repair strategies to restore damaged trust. Previous research has shown that expressing regret and providing explanations are effective strategies to rebuild human-agent trust. However, the role of various team dynamics, such as interdependence relationships, remains unexplored. The aim of this paper is to examine the influence of interdependence levels on the effectiveness of trust repair strategies. To investigate this, an experiment was conducted in a collaboration environment with an urban search and rescue mission. Two interdependence conditions were introduced to analyse their effect on trust and collaboration fluency. No significant evidence was found to support a relationship between interdependence and trust repair or collaboration fluency. However, as this study only considers two interdependence conditions, there is much more room for future work to explore further. This study can bring meaningful insights to design and facilitate agents that are more trustworthy in human-agent collaboration settings.

1 Introduction

Collaboration between humans and autonomous systems has been growing with rapidly developing technology. Systems can range from surgical robots that are physical and professional to virtual assistants like Siri which are more common in daily life. In human-agent teams (HAT), humans and autonomous agents work together to accomplish a common goal and often, they rely on each other to complete certain tasks to reach the goal. Interdependence relationships arise in situations where multiple parties engage in joint activities [1]. Here, joint activity refers to when a teammate depends on another teammate (and vice-versa) over a series of actions [1].

There are many factors that affect the success of HAT collaborations, including trust, observability, predictability, and directability [2,3]. Trust plays an important role in teamwork because the perceived trustworthiness of the agent determines how the human teammate interacts with it [4–6]. This paper aims to focus on human-agent trust in relation to interdependence as those two aspects are closely related [3].

Establishing and maintaining trust is essential for effective teamwork, and this is a challenging task as HAT situations are growing more complex. Unfortunately, in the real world, trust is often violated, for instance when an agent provides incorrect recommendations to the human [3]. This leads to damaging the human-agent trust. To compensate for the mistake, trust repair strategies are used by the agent, and it has been demonstrated that expressing regret and explaining the cause of the violation is an effective strategy [3,7]. The continuous re-adjustment of trust is vital to match the perceived trustworthiness to its actual trustworthiness and avoid 'overtrust' or 'undertrust' [8,9]. Therefore, it becomes important to understand the mechanisms behind trust repair strategies to aid trust calibration. Since trust is affected by various factors like interdependence, communication, task complexity, and explainability [10–12], investigating how this may be influenced by other situational factors is an interesting approach.

One of the variables that needs to be looked into is interdependence. This is because interdependence relationships serve as a mechanism for establishing trust [10]. However, the relationship between interdependence and trust repair strategies remains unexplored in previous research. By addressing this knowledge gap, future research can focus on refining trust repair strategies based on specific levels of interdependence and tailor interventions in order to optimize trust calibration in HAT.

There are multiple types of interdependence relationships depending on the components of the task, such as workflow or constraints [2, 13]. Complementary independence is an interdependence relationship that is often found in real-world HAT situations. Complementary independence emerges when the human and the agent collaborate with specific roles and tasks with complementing competencies. Because humans and agents have different capabilities, these relationships can improve efficiency and safety, in situations such as manufacturing assembly or search-and-rescue [14–16]. This paper will focus on comparing full independence and complementary independence to investigate how distinct roles and increased interdependence can affect trust evolution in HAT.

Another interesting factor to study in relation to interdependence is collaboration fluency. Collaboration fluency is how well a team is coordinated and how much the process is smooth and natural in a joint activity [17]. Measuring and improving collaboration fluency can lead to higher efficiency in human-agent teams, but more importantly, acceptance and confidence [17].

The research question of this paper is 'How does full independence and complementary independence in HAT influence trust repair and collaboration fluency?'. The relationship between interdependence and trust repair will be investigated by testing the hypothesis '*the trust repair strategy of expressing regret and explaining why the trust violation occurred will be more effective if the level of interdependence is higher (complementary independence)*'. In addition to trust, the paper will also look into the potential effect of interdependence on the collaboration fluency of the team. To answer the research question, a user study was conducted in a simulated environment, involving the collaboration between human participants and a virtual robot in a search and rescue (SAR) mission.

This research paper is structured as follows. In section 2, background information and relevant literature are introduced. Then, the methodology, including the experimental setup, is explained in section 3. Section 4 reflects on the ethical aspects of the research. Then the results of the experiment and the analysis are presented in section 5. In section 6, the contributions of the results are discussed, followed by limitations and suggestions for future work. Finally, section 7 provides the conclusion of the paper.

2 Background

2.1 Interdependence

Interdependence can be seen as the interconnections of teammates within the tasks [18]. As human-agent trust evolves over a series of interactions [19, 20], interdependence is also deeply intertwined with trust. Relational trust is formed and maintained through interdependence relationships [10]. The degree of interdependence can affect the level of trust, and even shape the form that trust takes [21].

Johnson et al. [22] identify two approaches to interdependence levels, *inter-activity dependence* and *intra-activity interdependence*. In inter-activity dependence, the types of interdependence are in terms of output in independent activities. Thompson's [23] types fall under this, where he suggested 3 types of interdependence; pooled, sequential, reciprocal [22]. The level of interdependence increases respectively [24]. Pooled interdependence is when all participants can execute tasks independently. Under sequential interdependence, tasks are performed in a certain order by different participants with different roles [25]. In reciprocal interdependence, they both require participants to work in turn, in a bidirectional manner. These interdependencies are also referred to as 'task interdependence' in some literature [13, 24, 25]. Inter-activity dependence is helpful to analyse the interactions and team dynamics at a higher level with respect to the final goal or product.

On the other hand, intra-activity interdependence pays attention to the atomic actions within a joint activity. Johnson et al. [2] identify two types of such interdependence, required and opportunistic. Required (hard) interdependency is when collaboration is required to accomplish the task while with opportunistic (soft) interdependency it is optional to increase efficiency, effectiveness or reliability. These classifications capture higher degrees of interdependence and nuanced interactions.

With the following types in consideration, full independence classifies as pooled interdependence. Complementary independence can fall under sequential interdependence or reciprocal interdependence depending on the collaboration environment. In the experimental setup of this paper, it is sequential as it does not involve bidirectional workflow in completing a task (see section 3.5). The inclusion of complementary and full independence conditions in this study brings forth precious insights. By initially examining the lowest two levels, pooled and sequential, the contrasting absence and presence of interdependence can be effectively explored. Additionally, the consideration of complementary independence, a relationship commonly observed in real-world scenarios [14–16], provides valuable context for understanding trust in such interdependence dynamics. Therefore, this study can serve as a promising starting point for investigating the relationship between interdependence and trust repair.

2.2 Trust and Trust Repair Strategies

Trust is vital in situations or relationships that has risk, uncertainty, or interdependence [26, 27]. This includes teamwork settings regardless of teammates, human or agent. There are diverse definitions of trust across disciplines with multidimensional models [27–29]. The definition by Mayer et al. is the most widely accepted trust definition, where the willingness to vulnerability was introduced as a key element [10, 21, 26, 29]. For the context of the paper, the definition of human-agent trust is adopted as " the human's willingness to make oneself vulnerable and to act on the agent's recommendations and decisions in the pursuit of some benefit, with the expectation that the agent will help achieve their common goal in an uncertain context" [3, p.30]. Having a high level of human-agent trust will lead to increased efficiency in tasks and ultimately better team performance [3].

Trust evolves over time, and it can strengthen or weaken based on the interaction. When the action of a party was untrustworthy, trust is violated. Competence-based trust violation is when the violation has to do with their competence, such as making errors in judgements while integrity-based violations relate to their integrity such as willingness or principles [29]. To compensate for the trust violation, the party may try to repair the trust using various strategies.

Prior research has examined the effectiveness of different combinations of trust-repair strategies in HAT collaboration environments. The most common strategies in human-agent or human-robot research include apologies, explanations, denial and promises [7]. Sebo et al. [30] demonstrated that apologies lead to higher trust compared to denial in a competence violation. In addition, situational factors like timing [30–32], severity and type of the violation [30, 33, 34] have been found to affect trust repair.

This paper mainly builds on the research by Kox et al. [3], where they established that expressing regret and explaining is an effective trust repair strategy for competence-based trust violations in HAT. In their task environment, an agent in robotic embodiment advised the human (participant) to seek shelter upon detecting danger. The advice was not always correct, and if there was an error, the agent used trust repair strategies, and the effectiveness was measured through questionnaires evaluating trust levels. It was shown that the damaged trust was significantly repaired when the apology included an expression of regret and that the effect was stronger with an explanation of why the trust violation occurred. These findings provide the foundation for further study into the dynamics of trust and trust repair strategies.

This study dives into unexplored situational factors in order to broaden the present understanding of trust repair strategies in HAT settings. Specifically, the focus is placed on interdependence, an essential aspect of human-agent interaction and trust dynamics [10]. This study seeks to fill the knowledge gap by examining the potential influence of interdependent relationships on the efficiency of trust repair strategies.

2.3 Collaboration Fluency

For success in a human-agent team setting, reaching a high level of coordination is crucial. This can include timings, effective and clear communication, the efficiency of tasks and how natural the whole process feels. This is measured through collaboration fluency. Striving for a high collaboration fluency not only increases the efficiency of tasks but also helps humans to accept their artificial teammate [17]. Fluency and efficiency are closely related, but they are not in-

Figure 1: Screenshot of the map of the environment in 'God' mode. This view with all of the victims and obstacles is only visible in the 'God' mode and is not visible while playing unless within the visible range.

Figure 2: Screenshot of the chat functionality in 'Human' mode in the start of the game.

terchangeable. With the growing need for human-agent collaboration, interest in collaboration fluency metrics has also grown. This research will measure collaboration fluency using both objective and subjective measures, drawing primarily from insights from G. Hoffman's research [17].

3 Methodology

3.1 Design

To test the hypothesis, an experiment was conducted in a simulated collaboration environment between a human and an agent. The experiment employed a 3X2 mixed design. The within-subject factor was the time the trust levels were measured, which occurred three times for each participant in distinct situations. These situations are: prior to trust violation [T1], after trust violation and repair strategy [T2], and after trust recovery [T3]. On the other hand, the two interdependence conditions were manipulated as a between-subject variable. These are full independence and complementary independence. The main dependent variables were trust and collaboration fluency.

3.2 Participants

A total of 30 participants took part in the study, with a majority being students from Delft University of Technology. The recruitment process involved selecting 15 participants for each interdependence condition. To minimize potential confounding factors, demographic information (age, gender, education, and gaming experience) was collected after obtaining consent from the participants. The participants consisted of 15 females and 15 males, 27 of them aged between 18 and 24 years, and the remaining three were aged between 25 and 34. In relation to education, one participant did some high school without obtaining a diploma, 24 participants were high school graduates, and five participants obtained a Bachelor's degree. In terms of gaming experience, three participants had no gaming experience, eight of them had little, five had moderate, seven had considerate and seven had a lot of gaming experience. The demographic information was used to distribute the demographics as evenly as possible while recruiting participants and allocating the condition.

Furthermore, statistical tests were conducted to ensure that the participants were evenly distributed. The Chi-square test of homogeneity was used to show that there were no significant differences for gender ($X^2 = 0.370$, $p = 0.543$). For age, education and gaming experience, Kruskal-Wallis was conducted. Results showed that the interdependence conditions were homogeneous in regards to age ($p = 0.550$), education ($p = 0.952$) and gaming experience ($p = 0.898$). Based on the analysis, it was deemed suitable to exclude the impact of demographic factors in the experiment and proceed accordingly.

3.3 Hardware and Software

The experiments were run on laptops with the collaboration environment installed and ready. Due to time constraints, the experiments were conducted in parallel with several peers, so it was not feasible to control all variables using a single laptop. However, measures were taken to ensure that the software operated as intended without any latency issues on all devices. To facilitate the research on human-agent teaming, a dedicated Python package called the Human-Agent Teaming Rapid Experimentation[1] (MATRX). This software enabled the creation of a simulated search and rescue task that was suitable for the experiment.

3.4 The Collaboration Environment

The environment was implemented digitally in a 2D world, on MATRX. In the environment, the human and the agent (RescueBot) had to search and rescue as many victims within

[1] https://matrx-software.com

Type of message	Message from the agent
Advice T1	I have detected extreme rain arriving soon and predict it will cause new floods, so I advise you to take shelter in one of the areas as soon as possible and until the rain is over
Feedback T1	My advice was correct, that weather was extreme! If you had not taken shelter, you would have lost important mission time due to injuries and 10 points of our score.
Advice T2	I have detected light rain arriving soon but predict it will cause no floods, so I advise you to continue searching and rescuing victims.
Repair T2	My advice was wrong. The amount of rain was heavy instead of light and because of that my flood prediction was incorrect. I am really sorry.
Advice T3	I have detected extreme rain arriving soon and predict it will cause new floods, so I advise you to take shelter in one of the areas as soon as possible and until the rain is over
Feedback T3	My advice was correct, that weather was extreme! If you had not taken shelter, you would have lost important mission time due to injuries and 10 points of our score.

Table 1: Overview of messages about the weather throughout the round

10 minutes. There were 4 mildly injured victims and 4 critically injured in the environment. They could be rescued by carrying the victim to their allocated drop zone. Additionally, an obstacle (stone, rock or tree) might have been blocking the entrance of the areas, and they needed to be removed to enter the area. The map of the environment is shown in Figure 1. The task had a time limit of 10 minutes to encourage the participants to carry out the tasks as efficiently as possible.

The agent and the human were able to communicate through limited messages, as visible in Figure 2. By clicking the buttons with messages written on them, the human was able to send which area they are going to search in, which victims they found in which area and whether they will pick up certain victims. The human could call for help to remove obstacles depending on the interdependence condition. The agent informed the human of their actions, including where it is heading, and which obstacles or victims it has found. Furthermore, it sent updates regarding the round's progress, including a list of located and rescued victims or the remaining time. Unless directed otherwise by the human (e.g. Help remove an obstacle), the agent searched the rooms automatically. However, it did not make decisions on active tasks by itself. Whenever it found a victim or obstacle, it notified the human about it and waited for the human to decide whether to do the task now or continue. These communications were also made through the chat.

Additionally, to simulate similar trust violations as the experience in the paper by Kox et al. [3], extreme weather was introduced instead of enemies. It could either rain heavily or lightly in the environment, and the human was advised to avoid heavy rains. Heavy rain injures the human, leading to the human freezing for 10 seconds and losing 10 points. The presence and penalty of extreme weather were informed before the round. The agent warned the human about the rain in advance with a message and gave feedback after the rain stopped. When the agent was correct about the weather, it gave feedback by saying that their advice was correct and explained what would have happened if the human did not or did follow its advice. However, the agent also made wrong predictions. In this case, it carried out the trust repair strategy through an apology with regret and an explanation. The actual messages can be found in Table 1. This happened 3 times throughout the whole mission with a 2-minute interval, so at 2 minutes ([T1]), 4 minutes ([T2]) and 6 minutes ([T3]) from the start of the task. At [T1], the agent made a correct prediction about heavy rain. At [T2], the agent predicted light rain and advises to continue searching. However, this advice was wrong and it rained heavily. Lastly, at [T3], the agent predicted heavy rain and this was again correct. This introduced a reliable mechanism for trust violation, enabling the agent to effectively implement trust repair strategies and foster dynamic trust development.

3.5 Interdependence Conditions

Viewing a task as 'rescuing a victim', this often follows a sequence of active sub-tasks: removing an obstacle and carrying a victim. The interdependence relationships were established by modifying the abilities of the team members to perform these sub-tasks. With full independence, both the agent and the human could individually remove obstacles or carry victims. In complementary independence, the agent was solely responsible for obstacle removal, while the human solely focused on carrying victims. Among the two conditions, complementary independence has a higher level of reliance and interdependence between the team members.

It is worth noting that the complementary independence condition in the experiment mirrors common real-life applications of search and rescue (SAR) robotics. In SAR missions, the agents often play a crucial role as site surveillance operators, exploring hazardous environments, identifying potential dangers, and locating survivors to mitigate the risk of humans [16]. Within the scope of this experiment, we simulated this scenario by assigning obstacle removal to the agent and allowing the agent to inform the human regarding obstacles and victims found.

3.6 Measures

3.6.1 Trust

The effectiveness of the trust repair strategy was measured by measuring the trust level at different times during a round. The participant answered a questionnaire about trust after every time it rained, meaning it was measured three times in a round. To achieve this, the questionnaire for trust measurement introduced by R. Hoffman was utilized [35]. This questionnaire comprises eight questions that employ a 5-point Likert scale ranging from 'strongly disagree' to 'strongly agree'. It was designed to measure trust and reliance in autonomous systems [35]. Here, the reliance asks how much the participant is likely to follow the agent's advice [35]. This questionnaire was deemed suitable for the experiment as it not only assesses the level of trust but also asks about reliance on the system's recommendations, which is a critical aspect to consider in the context of varying levels of interdependence. Moreover, the questionnaire is a validated tool from previous studies, which enhances the reliability and validity of the results obtained.

3.6.2 Collaboration Fluency

A questionnaire comprising eight indicators of a subjective sense of fluency was employed to assess collaborative fluency [17]. This questionnaire evaluates various dimensions of collaborative fluency, being general fluency, robot contribution, commitment, and teammate traits [17]. This was done through a 7-point Likert scale from 'strongly disagree' to 'strongly agree'. The participants answered this questionnaire only once after the whole round. The questionnaire was chosen as it allows for the assessment of subjective perceptions of fluency experienced by participants during the collaboration. This approach is particularly valuable to provide insights about the nuanced aspects of collaboration that are not easily captured by objective measures alone.

3.6.3 Performance

Although it does not directly answer the research question, the performance of the team is also an important measure that is worth investigating, especially in relation to collaboration fluency. Three objective metrics were used in this regard: score, completeness and time taken to finish a round. These metrics serve as valuable indicators for evaluating the team's effectiveness and efficiency in achieving their objectives.

The score of the round was calculated automatically and recorded in the output logs. 3 points were awarded for mildly injured victims and 6 points for the critically injured. 10 points were deducted whenever the human was out in the rain. The maximum points attainable in a round was 36 points. This metric is interesting as it also incorporates advice acceptance, as it deducts points for being out in heavy rain.

Completeness refers to the degree to which a mission has been entirely achieved, specifically in terms of rescuing injured victims. On a scale from 0 to 1, it quantified the ratio of the total 8 injured victims who have been successfully rescued. This measure is effective to focus on the efficiency of the round, being how many victims were actually rescued, without considering the trust violations.

Lastly, the time taken for the round was also measured automatically with the tick system in MATRX, and recorded in the logs. As mentioned earlier, there was a time limit in a round, so the max value for this metric was 600 seconds.

3.7 Procedure

The experiment took place in person, where the participants carried out the task and answered the questionnaires on the provided laptop. Before the actual experiment, the participants read and signed the consent form and filled in their demographic information. Then they played the tutorial, where the task, controls, messaging and the overall system is introduced. In the main round, the participants were given 10 minutes to rescue as many victims as possible. To help them with the task interdependencies, a printed cheat sheet with intuitive images and explanations was accessible at all times. After each rain, which occurs 3 times in a round, the task is stopped automatically. Then the participants were directed to the questionnaire to answer the questionnaire to record trust. After the round, the collaboration fluency questionnaire was presented to evaluate fluency of the whole round.

4 Responsible Research

4.1 Ethics

The user study described in this paper received approval from the Human Research Ethics Committee (HREC)[2] at TU Delft. In accordance with the HREC checklist, risks were identified prior to conducting the experiments, and measures to mitigate these risks were planned and implemented. Notably, all participants involved in the study were voluntary and did not possess any vulnerability. Prior to the experiment, participants were provided with a consent form that explicitly outlined the experiment's objectives, associated risks, and data storage procedures. Through this process, participants were fully informed and provided consent for their involvement, as well as for the storage of anonymized data. Participants were also made aware of their right to withdraw from the experiment at any point.

One prominent risk identified during the study pertained to the potential for re-identification, given the collection of demographic information. However, it is crucial to note that the questions posed did not seek specific or personally identifiable details. Rather, they focused on obtaining general information such as gender, education level, age (within a six-year range), and gaming experience. Moreover, the collected data underwent anonymization processes, making it highly improbable to re-identify participants solely based on the demographic information gathered.

4.2 Reproducibility

The reproducibility of the experiment can be established based on several factors. Firstly, a comprehensive description of the experimental conditions and procedures is provided in Section 3 of the study. The questionnaire utilized in the experiment is referenced and accessible for further examination. Furthermore, the complete codebase, which includes the MATRX collaboration environment, tutorial, all

[2] https://www.tudelft.nl/en/about-tu-delft/strategy/integrity-policy/human-research-ethics

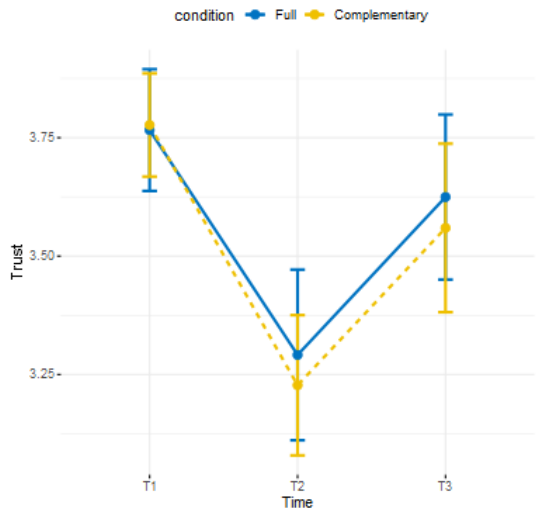

Figure 3: Development of trust levels over time

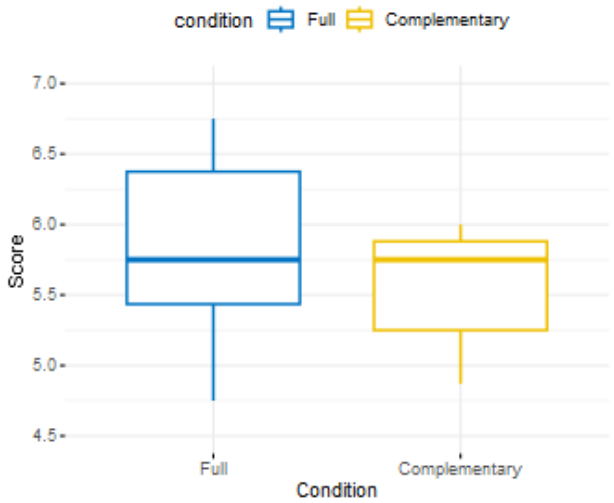

Figure 4: Box plot for collaboration fluency

the relevant interdependence conditions, and the code for automatic logging of objective measures, is openly available on GitHub[3]. By engaging with the environment and participating in a round, all actions and objective measures are automatically recorded in the logs. This allows readers to collect the objective measures in the same manner as performed during the experiment and facilitates inspection of the data collection process. Additionally, this functionality minimizes the potential for manipulating results. Such transparency ensures that the experiment's data remains open to reanalysis at any given time. It is important to note that the data was collected without cherry-picking or neglecting any observations, ensuring unbiased conclusions. The reader should be able to replicate the experiment and verify the results by carrying out the experiment with similar demographics and following the provided procedures.

5 Results

5.1 Trust

As shown in Figure 3, there is a general trend in how the trust levels evolve over time. At [T1], the mean in full independence is 3.77 (sd = 0.498) and the mean in complementary independence is 3.78 (sd = 0.422). At [T2], the mean is 3.29 (sd = 0.698) in full and 3.23 (sd = 0.573) in complementary. Lastly, at [T3], the mean is 3.62 (sd = 0.675) in full and 3.56 (sd = 0.689) in complementary. In general, trust in [T1] is relatively high, and it decreases after the trust violation in [T2] and recovers by [T3]. The graph gives some intuitive insights, but further analysis should be conducted to check if these are statistically significant.

The data-set met the assumptions for a two-way mixed ANOVA (no extreme outliers, normality, assumption of sphericity and homogeneity of covariances) except for the homogeneity of variances checked through Levene's test (p = 0.002 at [T3]). Therefore, a robust ANOVA was conducted as an alternative test. The two factors were condition (Full, Complementary) and time ([T1], [T2], [T3]).

The test did not reveal any interaction effect between condition and time. However, there was a statistically significant main effect of time (p = 0.005) on trust. The Friedman test was performed as a post-hoc to detect a small effect (p = 0.002, Kendall W = 0.2). Wilcoxon signed-rank tests with the Bonferroni adjusted p-value were used for pairwise comparisons. Results showed that there were significant differences between [T1-T2] (p.adj = 0.005) and [T2-T3] (p.adj = 0.008) but not in [T1-T3] (p.adj = 0.486). Overall, the analysis showed that there are no significant differences between the interdependence conditions. In contrast, it supports a significant trust violation in [T2] and a significant recovery [T3] in both conditions. The insignificant difference between [T1-T3] indicates that the trust increase in [T2-T3] was sufficient to recover back to the initial trust level in [T1].

5.2 Collaboration Fluency

As the data set for collaboration fluency did not have a normal distribution, the Mann-Whitney test was used as a non-parametric alternative to examine the differences in fluency across the conditions. However, there were no statistically significant differences (p = 0.307). This can be also seen in Figure 4, where the means are at a similar level in the two conditions.

5.3 Performance

For the metrics for performance, none of the three (score, completeness, time taken) met the assumptions for the independent samples t-test, so a Mann-Whitney test was conducted. For the scores, the mean for full independence was 22.7 (sd = 10.8) while the mean for complementary independence was 18.9 (sd = 15). The Mann-Whitney test revealed that there were no significant differences (p = 0.538).

For completeness, the mean was 0.975 (sd = 0.070) for full independence and 0.827 (sd = 0.229) for complementary independence. There were statistically significant differences

[3]https://github.com/mawakeb/CSE3000-2023-trust-repair

between the conditions (p = 0.009). Vargha and Delaney's A showed that the effect size is large (0.744), meaning that the completeness in the full independence dominates the completeness in the complementary condition.

For time taken, the mean was 494.4 seconds (sd = 17.0) for full independence and 586.4 (sd = 25.5) for complementary independence. The Mann-Whitney test revealed that there is a highly significant difference ($p < 0.001$). Vargha and Delaney's A showed that the effect size is large (0.116).

6 Discussion

6.1 Trust

As shown in section 5.1, there is a significant trust decrease between [T1] and [T2] and a significant recovery between [T2] and [T3]. However, no evidence was found to support the differences in the effectiveness of the repair strategy across the interdependence conditions. Therefore, the hypothesis '*the trust repair strategy of expressing regret and explaining why the trust violation occurred will be more effective if the level of interdependence is higher (complementary independence)*' was not supported.

Previous research has mixed results regarding the impact of interdependence on HAT collaborations, with some studies suggesting positive effects [36, 37] while others demonstrated that higher interdependence leads to decreased trust [12]. However, this research does not align with either perspective, as no evidence was found to support any significant relationships between interdependence and trust. One possible explanation may be that factors other than interdependence play a more significant role in trust repair. Research has extensively explored the influence of factors such as anthropomorphism [38, 39], type of repair strategy [3, 30, 38], type of trust violation [30, 33, 34], and timing of the repair strategy [30–32]. However, there remains a gap in understanding the impact of task-related factors, such as interdependence or task difficulty [7].

Existing research encompasses a wide range of tasks, including military simulations, emergency scenarios, driving games, and manufacturing tasks [3, 7, 36]. Consequently, the levels of interdependence within these studies vary considerably, which is undesirable given the linkage between interdependence and the establishment and maintenance of trust [10, 12]. To draw meaningful conclusions, it is crucial to investigate the nature of tasks and their effects.

While this experiment did not yield significant findings regarding interdependence, we hypothesize that further relationships may emerge by incorporating a broader range of interdependence conditions with greater differences. Therefore, future studies should consider task-related factors like interdependence to gain a more comprehensive understanding of trust repair strategies within HATs. These insights hold implications for the design and tailoring of trust repair strategies in HATs, ultimately enhancing their overall success and functionality in various domains.

6.2 Collaboration Fluency

The experiment did not reveal any correlations between interdependence and collaboration fluency, as there were no significant differences between the two conditions. Previous research suggested that sequential interdependence has the lowest subjective fluency out of pooled, sequential and reciprocal interdependence [13]. This was in terms of human-idle time and the perception of the robot and participants reported that they felt a lack of freedom in sequential interdependence. Our experiment did not find evidence to support the previous study. A reason could be the differences in task complexity and autonomy within the same conditions. If the lack of individual autonomy [13] played a huge role to lower subjective fluency in sequential interdependence, this was not the case in this experiment. Our collaboration environment gave a larger range of possible actions while the agent was working on the task. For further research to substantiate this claim, it would be critical to take the autonomy of the human into account while making such comparisons. It would be also interesting to include reciprocal interdependence to compare with the current conditions.

Although job performance does not directly equate to collaboration fluency, the performance metrics in this paper can still serve as indicators of collaboration fluency [17,40]. Time taken is often used as an objective metric for collaboration fluency [17, 30] and higher fluency fosters higher job performance [41, 42]. However, we speculate that the significant effects found in completeness and time taken were due to the nature of the different conditions in this experiment. It is much slower to do the task in sequential interdependence because they often have to wait for each other to do certain tasks to proceed, inhibiting the performance of the team. In the future, it may be useful to implement different interdependence conditions does not affect the efficiency so much to draw more meaningful conclusions about collaboration fluency with these metrics. For instance, sequential interdependence could be implemented in a collaboration environment where the teammates would not have to wait for their partner to arrive from the other side of the map. Overall, this experiment did not find evidence that collaboration fluency is influenced by interdependence levels.

6.3 Limitations and Future Work

Due to the nature of the project, there are some limitations. One of them is the relatively small number of participants. The current experiment had 30 participants in total, with only 15 participants for each condition. A larger sample would strengthen the statistical power of the results and possibly reveal new relations. Future research can be made with a larger and more diverse pool of participants to enhance the external validity and generalizability of the study.

Another limitation is that the trust violation and the repair strategy might have gone unnoticed by the participants. Despite our efforts to enhance their awareness through sound effects and bright colors in the messages, there were instances where participants were too engaged in their tasks and failed to perceive the intended messages. This could have been improved with audio actually reading the warnings and the message containing the trust repair strategy so that participants can understand it without looking at the messages. Additionally, some participants found themselves in the shelter coincidentally while rescuing victims, which could have made the

trust violation (incorrect advice) less significant. This was partially recorded through an extra question asking if their actions were based on the agent's advice. However, due to time constraints, these cases could not have been eliminated and there has not been a thorough analysis of whether this actually affected the perceived trust violation. Investigating this in the future can help to substantiate the results of this experiment or design another experiment to prevent this. These deviations from the original experiment conducted by Kox [3] should be acknowledged, as in their study participants always experienced apparent trust violations resulting from incorrect advice. In contrast, our experiment provided participants with the choice to follow the advice or not, which led to more complex scenarios like not having enough time to take shelter although they wanted to follow their advice or being in a shelter doing some other task. In hindsight, a potential resolution to this limitation could have been creating a separate shelter exclusively designed to protect against rain, ensuring that participants would be more attentive to the warnings and eliminating the possibility of being in a shelter coincidentally.

Furthermore, it is important to recognize that the types of interdependence considered in the study are limited. Both full independence and complementary independence have relatively low levels of interdependence. Although the current experiment did not find evidence supporting the influence of interdependence on trust repair and collaboration fluency, it is essential to consider future research with other interdependence conditions to unveil potential relationships. For instance, reciprocal interdependence can be also included to compare all 3 types of interdependence that Thompson identified [23]. Alternatively, investigating the intra-activity interdependencies, including both required and opportunistic interdependence, can shed light on the intricate atomic relationships, moving beyond a higher-level analysis of the collaboration.

In addition to the primary focus on trust and collaboration fluency, it would be interesting to explore and analyze other factors measured in the experiment. There were other measurements that were logged automatically in the environment, such as the ratio of idle time and the number of messages exchanged. Among these factors, one particularly interesting one to investigate would be advice acceptance. This was not only measured through the logs (a true/false value for which they took shelter during the rain) but also through an extra question in the questionnaire to ask whether they considered the agent's advice while making the decision. Advice acceptance is a valuable indicator to human-agent trust, but its analysis had to be eliminated due to its complexity and time constraints. Furthermore, advice acceptance was measured in the study by Kox et al. [3], so a comparison can be made with their results. Overall, these objective variables, which examine the participants' actual behaviors rather than their perceptions, have the potential to reveal fresh insights to understand how individuals react to trust violations and repairs.

7 Conclusion

The aim of this research was to investigate the influence of interdependence on the effectiveness of trust repair strategies and collaboration fluency in human-agent teams (HAT). In particular, the study focused on exploring the influence of two distinct forms of interdependence: complementary independence (sequential) and full independence (pooled), wherein complementary independence had a higher degree of interdependence relative to the latter condition. To answer the research question, a user study was conducted in a simulated collaboration environment, with a search and rescue (SAR) mission scenario. Upon trust violation, the agent gave an apology and an explanation as a trust repair strategy.

The findings of this investigation unveiled several noteworthy insights. First and foremost, the analysis revealed no statistically significant differences in trust levels across the interdependence conditions, indicating that interdependence did not significantly impact the effectiveness of trust repair strategies and collaboration fluency in this setting. However, it should be highlighted that the study did observe significant variations in trust violation and subsequent recovery between the measured time points. Furthermore, the study did not find substantial evidence regarding the influence of interdependence on collaboration fluency. It is important to recognize that this investigation has certain limitations, primarily the fact that only two types of interdependence were considered. Consequently, there exists ample scope for future research to expand the spectrum of interdependence conditions, thereby facilitating a more comprehensive understanding of the relationship.

Overall, this paper offers a scientific methodology for investigating the impact of interdependence on trust repair, while also serving as a foundational stepping stone towards the optimization of trust repair strategies. The implications of this research extend to developing agents that are perceived as trustworthy by human counterparts, even in scenarios where failures are inevitable. Moving forward, these findings have the potential to contribute to the field, ultimately fostering the creation of more reliable and effective human-agent teams in a variety of domains.

References

[1] Matthew Johnson, Jeffrey Bradshaw, Paul J. Feltovich, Catholijn Jonker, Birna Riemsdijk, and Maarten Sierhuis. The fundamental principle of coactive design: Interdependence must shape autonomy. In *Coordination, Organizations, Institutions, and Norms in Agent Systems VI*, volume 6541, pages 172–191, 01 2010.

[2] Matthew P. Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna Van Riemsdijk, and Maarten Sierhuis. Coactive Design: Designing Support for Interdependence in Joint Activity. *Journal of human-robot interaction*, 3(1):43, 2 2014.

[3] Esther S. Kox, Johanna H. Kerstholt, Tom A. Hueting, and P. De Vries. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems*, 35(2), 6 2021.

[4] G. H. Klien, David D. Woods, John D. Bradshaw, Robert M. Hoffman, and Paul J. Feltovich. Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intelligent Systems*, 19(06):91–95, 11 2004.

[5] Da-jung Kim and Youn-kyung Lim. Co-performing agent: Design for building user-agent partnership in learning and adaptive services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery.

[6] Frances S. Grodzinsky, Kurt Miller, and Matthias Wolf. Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?". *Ethics and Information Technology*, 13(1):17–27, 12 2010.

[7] Connor Esterwood and Lionel P. Robert. A literature review of trust repair in hri. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1641–1646, 2022.

[8] E.S. Kox, L.B. Siegling, and Johanna Helena Kerstholt. Trust Development in Military and Civilian Human–Agent Teams: The Effect of Social-Cognitive Recovery Strategies. *International Journal of Social Robotics*, 14(5):1323–1338, 4 2022.

[9] Carolina Centeio Jorge, Myrthe L. Tielman, and Catholijn M. Jonker. Artificial trust as a tool in human-ai teams. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1155–1157, 2022.

[10] Matthew P. Johnson and Jeffrey M. Bradshaw. *The role of interdependence in trust*. Elsevier BV, 1 2021.

[11] Peter A. Hancock, Deborah L. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. De Visser, and Raja Parasuraman. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, 53(5):517–527, 10 2011.

[12] Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. *Frontiers in Robotics and AI*, 9, 9 2022.

[13] Fangyun Zhao, Curt Henrichs, and Bilge Mutlu. Task interdependence in human-robot teaming. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1143–1149, 2020.

[14] Selma Musić and Sandra Hirche. Control sharing in human-robot team interaction. *Annual Reviews in Control*, 44:342–354, 2017.

[15] L. Wang, R. Gao, J. Váncza, J. Krüger, X.V. Wang, S. Makris, and G. Chryssolouris. Symbiotic human-robot collaborative assembly. *CIRP Annals*, 68(2):701–726, 2019.

[16] Teng H. Chan, James Kusuma, Kian Tan, Emmanuel Tang, Wei J. Ang, Jin Y. Tan, Samuel Cheong, Hoan-Nghia Ho, Benson Kuan, Muhammad Shalihan Bin Othman, Ran Liu, Gim Soh, Chau Yuen, U-Xuan Tan, Lionel Heng, and Shaohui Foong. A robotic system of systems for human-robot collaboration in search and rescue operations. 06 2023.

[17] Guy Hoffman. Evaluating Fluency in Human–Robot Collaboration. *IEEE Transactions on Human-Machine Systems*, 49(3):209–218, 4 2019.

[18] Andrew H. Van de Ven, Ricky Leung, John P. Bechara, and Kangyong Sun. Changing organizational designs and performance frontiers. *Organization Science*, 23(4):1055–1076, 2012.

[19] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in Human-Agent Real-Time Collaboration. *Frontiers in Robotics and AI*, 8, 7 2021.

[20] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, page 164–168, New York, NY, USA, 2016. Association for Computing Machinery.

[21] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin F. Camerer. Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review*, 23(3):393–404, 7 1998.

[22] Matthew P. Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, Birna Van Riemsdijk, and Maarten Sierhuis. *The Fundamental Principle of Coactive Design: Interdependence Must Shape Autonomy*. Springer Science+Business Media, 1 2011.

[23] James D. Thompson. *Organizations in Action*. McGraw-Hill Companies, 1 1967.

[24] Ronal Singh, Liz Sonenberg, and Tim Miller. *Communication and Shared Mental Models for Teams Performing Interdependent Tasks*. Springer Science+Business Media, 5 2016.

[25] Richard Saavedra, P. Christopher Earley, and Linn Van Dyne. Complex interdependence in task-performing groups. *Journal of Applied Psychology*, 78(1):61–72, 2 1993.

[26] Roger Mayer, James Davis, and F. David Schoorman. An Integrative Model Of Organizational Trust. *Academy of Management Review*, 20(3):709–734, 7 1995.

[27] D. Harrison McKnight and Norman L. Chervany. What is Trust? A Conceptual Analysis and an Interdisciplinary Model. *AMCIS 2000 Proceedings*, 1 2000.

[28] John D. Lee and Katrina See. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1):50–80, 1 2004.

[29] Anthony J. Baker, Elizabeth J. Phillips, Daniel Ullman, and Joseph R. Keebler. Toward an Understanding of

Trust Repair in Human-Robot Interaction. *ACM transactions on interactive intelligent systems*, 8(4):1–30, 11 2018.

[30] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. "i don't believe you": Investigating the effects of robot trust violation and repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 57–65, 2019.

[31] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. *Timing Is Key for Robot Trust Repair*. Springer Science+Business Media, 1 2015.

[32] M. Nayyar and A. R. Wagner. When should a robot apologize? understanding how timing affects human-robot trust repair. In *Lecture Notes in Computer Science*, Chapter 26, pages 265–274. Springer International Publishing, 2018.

[33] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 507–513, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[34] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 3–12, New York, NY, USA, 2020. Association for Computing Machinery.

[35] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in computer science*, 5, 2 2023.

[36] Thomas O'neill, Nathan McNeese, Amy Barron, and Beau Schelble. Human-autonomy teaming: A review and analysis of the empirical literature. *Human Factors The Journal of the Human Factors and Ergonomics Society*, 64:1–35, 10 2020.

[37] James Walliser, Ewart de Visser, Eva Wiese, and Tyler Shaw. Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, 13:155534341986756, 08 2019.

[38] Tae-Nyun Kim and Hayeon Song. How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, 61:101595, 8 2021.

[39] Elizabeth Phillips, Kristin E. Schaefer, Deborah R. Billings, Florian Jentsch, and Peter A. Hancock. Human-animal teams as an analog for future human-robot teams: Influencing design and fostering trust. *J. Hum.-Robot Interact.*, 5(1):100–125, mar 2016.

[40] Guy Hoffman and Cynthia Breazeal. Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In *HRI '07: Proceedings of the ACM/IEEE international conference on Human-robot interaction*, HRI '07, page 1–8, New York, NY, USA, 2007. Association for Computing Machinery.

[41] Mateusz Paliga and Anita Pollak. Development and validation of the fluency in human-robot interaction scale. a two-wave study on three perspectives of fluency. *International Journal of Human-Computer Studies*, 155:102698, 07 2021.

[42] Luca Gualtieri, Erwin Rauch, and Renato Vidoni. Emerging research fields in safety and ergonomics in industrial collaborative robotics: A systematic literature review. *Robotics and Computer-Integrated Manufacturing*, 67:101998, 2021.