Agent Failure, Trust Repair, and Fluency in Human-AI Teams

Impact of Opportunistic Interdependence Relationship on Trust Violation, Trust Repair, and on Collaboration Fluency in a Human-Agent Team

Kanta Tanahashi

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: Kanta Tanahashi
Final project course: CSE3000 Research Project
Thesis committee: Myrthe Tielman, Ruben Verhagen, Ujuwal Gadiraju

Abstract

Nowadays, Human Autonomy Teams (HATs) are incorporated in many fields, where humans and autonomous agents work collaboratively to combine their capabilities with the ultimate goal of performing tasks more efficiently. In such environments, it is imperative to sustain a high level of trust between the agents as collaboration is not possible without mutual trust. Naturally, this implies that recovering trust following trust violation is also a crucial aspect of HATs. Moreover, besides team performance, the fluency of collaboration is another important factor to consider when evaluating the success of the teams. This paper aims to investigate the effect of opportunistic (soft) interdependence between the agents on trust violation, trust repair, and on collaboration fluency when compared against a baseline (complete independence) condition. In this paper, interdependence relationships refer to how the agents complement/combine each other's competence. The experimental results were obtained through a user study, using questionnaires and logged objective metrics. Our research found that teams with opportunistic interdependence relationships were significantly affected by trust violations compared to the baseline condition. Furthermore, although not as significant as the effect of trust violation, they also experienced a significant trust recovery during the tasks. Finally, the results of analyzing both subjective and objective fluency metrics did not give any significant result that indicates the difference in the level of collaboration fluency between the two conditions.

1 Introduction

Teams composed of humans and autonomous AI agents, so-called Human Autonomy Teams (HATs) are becoming increasingly common in many different segments. An example of this is Robonaut developed by NASA which is a humanoid robot that can assist astronauts to perform tasks in space [1]. There exists some level of interdependence relationships between the agents, where interdependence refers to the relationship between the agents in terms of the tasks they can perform. They combine and complement their capabilities based on their interdependence relationships, with the ultimate aim to improve task efficiency as a team. Furthermore, a collaboration between multiple agents implies that a sufficient level of trust needs to be sustained to maximize the benefits of working collaboratively [2, 3]. It is possible for the trust to be violated if, for example, one agent provides incorrect information or advice to the other agent. In such cases of trust violations, a trust repair strategy needs to be enforced to recover trust. Prior research found that expressing regret and providing explanations for the trust violation significantly helped recover human trust in the AI agent [4].

Another interesting factor to consider in HATs is collaborative fluency and it is a measurement of coordination and meshing of actions in a team [5]. Indeed, when evaluating the success of HATs, it is imperative not only to assess task efficiency but also to examine how smoothly the joint actions are performed from the perspective of human agents. This is evident from the research done by Hoffman and Breazeal, where subjects' sense of fluency was not always correspondent to their task efficiency [6].

Core concepts of HATs, such as interdependence, trust violation, trust repair, and collaboration fluency have been researched independently. For instance, many studies have looked into mitigating the loss of trust following trust violations and achieving effective trust repair [7–9]. In addition, some prior research has investigated the role of interdependence relationships on trust [10–12]. Despite these, there is no prior research done that investigates the role of interdependence on trust violation, trust repair, or on collaboration fluency in particular. This, together with an increasing demand for human-AI collaboration, makes it an interesting area to conduct further research. Consequently, this led our research group, in collaboration with AI*MAN Lab EEMCS [1], to fill out the research gap, investigating the effect of different interdependence relationships on trust violation, trust repair, and on collaboration fluency. Furthermore, this paper lays emphasis on opportunistic (soft) interdependence relationships. This was inspired by the claims made by Johnson et al. in which they state that many aspects of teamwork involve soft interdependence [13]. Additionally, they argue that their past studies show that the success of HATs is often determined by how well soft interdependencies are managed. Looking at the bigger picture, this research helps us gain insights into establishing HATs involving soft interdependence that can sustain a consistently high trust level amongst agents and achieve high fluency in collaboration from the perspective of human agents.

Formally, this report aims to answer the following research question: ”How do opportunistic interdependence relationships affect 1) the trust violation and trust repair and 2) collaboration fluency in a Human-AI team, when compared against a complete independence (baseline) condition?”.

This report is structured as follows. In Chapter 2, the background of this research, including concepts of trust, interdependence, and collaboration fluency, is elaborated. Chapter 3 discusses the methodology used for this research, and responsible research in terms of ethics and reproducibility of methodology is explained in Chapter 4. Chapter 5 presents obtained results, followed by a detailed discussion of the results in Chapter 6. Finally, the report ends with a conclusion in Chapter 7.

ChatGPT was used to rephrase some words/expressions in my sentences to construct more formal sentences and to improve writing flows in the ”Environment” and ”Procedure” subsections of the Methodology section

[1] https://www.tudelft.nl/ai/aiman-lab

2 Background

2.1 Trust Violations and Trust Repair

Trust has been found to play a crucial role in Human Autonomy Teams based on past studies. For instance, McNeese et al. found that the level of human trust in the autonomous agent is lower in low-performing teams when compared to both medium and high-performing teams [2]. Additionally, Burke et al. claim that a high level of trust in the teams leads to effective mutual performance monitoring [3]. (i.e. "the ability to develop common understandings of the team environment and apply appropriate task strategies to accurately monitor teammate performance")

Multiple definitions of trust have been proposed in the past. For example, Lee and See claim that trust can be defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [14]. Furthermore, Kox et al. define trust as "the human's willingness to make oneself vulnerable and to act on the agent's decisions and recommendations in the pursuit of some benefit, with the expectation that the agent will help achieve their common goal in an uncertain context" [7]. Both definitions are similar in the sense that trust is directional from one agent to another and is considered necessary to complement vulnerabilities in the team. In this case, the word 'directional' is used to imply that trust only exists under the presence of the target agent. (i.e. trustee)

In any HATs, trust violations are sometimes inevitable, let it be due to a lack of competence or lack of integrity in actions. It is therefore imperative to understand ways in which the damage of trust violations can be mitigated. A past study has found that communicating uncertainty in the advice of the agent helps minimize the loss of trust following trust violations [7].

Similarly, following trust violations, lost trust needs to be recovered. This is often achieved by implementing an effective trust repair strategy. Multiple research has been conducted, which revealed that the effective trust repair strategy is dependent on the type of trust violation. One study found that providing an apology was an effective trust repair strategy in case of competence-based violations whereas denials were found to provide an effective response to integrity-based violations [8]. Focusing more on apology as a means of trust repair, another study revealed that providing an apology with an internal attribution in case of competence-related trust violations was effective, whereas for integrity-related trust violations, apologizing with external attributions resulted in a more significant trust repair [9].

Ultimately, some research has been done on effective trust repair in human-robot teaming specifically. Kox et al. found that expressing regret and adding explanations for trust violations significantly increased the effectiveness of a trust repair strategy [4]. Furthermore, Baker et al. have put together five recommendations for advancing research in the area of human-robot trust repair, which includes investigating the role of culture in human-agent trust repair [15].

2.2 Interdependence Relationships

Human Autonomous Teams have some level of interdependence between human agents and AI agents, ranging from complete independence to required (hard) interdependence. Johnson et al. define interdependence as "the set of complementary relationships that two or more parties rely on to manage required (hard) or opportunistic (soft) dependencies in joint activity" [13]. Furthermore, he claims that opportunistic interdependence "arises from recognizing opportunities to be more effective, more efficient, or more robust by working jointly", suggesting that collaboration is optional and not strictly required. In the bigger picture, the opportunistic interdependence relationship can place itself somewhere in the middle in terms of the level of interdependence. Complete independence entails that agents can only perform tasks independently, suggesting the lowest level of interdependence. On the other end of the interdependence level spectrum, there is hard interdependence, which strictly requires collaboration to complete tasks.

Multiple studies have been conducted which highlight the importance of interdependence relationships in HATs. For example, it was found that interdependence causes significant behavioral conformity when having a computer as a teammate [16]. Another interesting research has found that simply increasing the autonomy level of an AI agent without accounting for the interdependence in a human-robot collaboration does not always lead to better performance [17]. This suggests the importance of correctly managing interdependence relationships between the agents. Indeed, Hoffman et al. claim that the focus around autonomous system development has already been shifted from simply trying to create independent agents by making them more autonomous, to also striving to make them competent in performing interdependent joint activities with humans [18].

Despite the presence of multiple levels of interdependence, the effective management of opportunistic interdependence is shown to significantly impact the performance of HATs in particular. In fact, Johnson et al. believe that despite the fact that much of the robotics work nowadays deals with hard constraints between the agents, the presence of an opportunistic interdependence relationship is necessary to achieve true teamwork [13]. Indeed, based on their past studies, they claim that extent to which opportunistic interdependencies are managed well tends to determine the success of the team. This highlights the importance of investigating opportunistic interdependence in particular, to understand the optimal way in which soft interdependence can be incorporated into HATs, leading to achieving greater performance and teamwork.

Finally, it is worth mentioning that some past research has been done on the role of interdependence relationships on trust. One research has found the general role of interdependence relationships, stating that they "support the active and continuous exploration of trust between a trustor and a trustee to ensure trustor assessments are appropriate for achieving the best outcomes possible" [10]. Furthermore, studies on the relationship between interdependence and trust have produced contradictory results in the past. While one research has found that having a team structure in a team consisting of a participant and another agent increases trust when compared to having no team structure [12], another research has found that a high level of interdependence significantly decreases the trust level [11].

2.3 Collaboration Fluency

Even though many believe that the main purpose of collaboration between humans and autonomous agents is to increase overall task efficiency, the success of HATs cannot simply be measured through performance. Indeed, one can argue that the measurement of coordination and meshing of actions in a team, so-called collaborative fluency is an important factor to be considered. The research done on human-robot collaboration by Hoffman supports this argument [6]. In the research, subjects were asked to assemble a cart made of different components in a simulated factory setting, together with a robot. The study showed a significant difference in perceived fluency of collaboration while the measured task efficiency was similar across participants. Consequently, this implies that perceived fluency in collaboration needs to be separately measured along with task efficiency, in order to measure the level of success in HATs.

Formally, Hoffman defines collaborative fluency as ”the coordinated meshing of joint activities between members of a well-synchronized team” [5]. From this definition, we can conclude that collaborative fluency concerns the quality of the collaboration process rather than the team outcome achieved through collaboration.

Furthermore, he discusses two types of fluency metrics for human-robot collaboration in his paper; subjective fluency metrics and objective fluency metrics.

Subjective metrics measure the human perception of the fluency of interaction and related qualities of the robot [5]. More specifically, he refers to subjective fluency metrics to include direct measures of fluency of a collaboration and downstream outcomes of the perceived fluency. Examples of the downstream outcomes include robot relative contribution and positive teammate traits. In addition, the paper mentions how subjective fluency metrics are often measured through the use of questionnaires to rate agreement with fluency notions. Section 3.7.2 discusses how the subjective fluency metric items from his paper were adopted in our research.

On the other hand, objective fluency metrics aim to utilize objective measurements and tie them to perceived fluency [5]. In his paper, he mainly introduces the following four objective metrics: human idle time, robot idle time, concurrent activity, and functional delay. He further claims that more appropriate metrics may exist depending on the collaborative scenarios. More information on objective metrics utilized in our research can be found in section 3.7.2

3 Methodology

3.1 Design

A three-by-two mixed design was used to measure the trust level. That is, the level of human trust in an autonomous AI agent was measured in three different timestamps (prior to trust violation, after trust violation and repair strategy, and after trust recovery) for two different interdependence relationships. (opportunistic and baseline) Time was a within-participant factor, and type of interdependence was a between-participant factor. Collaboration fluency was measured once in each user study.

3.2 Participants

In total, 30 participants were recruited. The participants consisted mostly of students from TU Delft, however, some were also recruited outside the university. Participants were assigned to one of the two interdependence conditions, resulting in 15 participants being recruited for each condition. Furthermore, demographic information, including age, gender, education, and gaming experience, was collected and attempted to be balanced. The Chi-Square Test of Homogeneity was performed for gender while for others, the Kruskal-Wallis test was performed to check for a significant difference in demographics between the two conditions. The obtained p-values were all greater than 0.05 (age: $p = 0.317$, gender: $p = 0.111$, education: $p = 1$, gaming experience: $p = 0.311$), indicating that there is no statistically significant demographic difference between the two conditions.

3.3 Hardware and Software

In this project, the human-agent teaming rapid experimentation software package (MATRX[2]) was used to simulate the experimental environment. Participants ran the simulated environment on Google Chrome or Mozilla Firefox through their/conductor’s laptop.

3.4 Environment

The MATRX environment simulated an environment where a human agent and an autonomous AI agent (ResuceBot) collaborate together to complete a search and rescue task. The objective of the simulation was to rescue as many victims as possible by carrying them to a drop zone within a given time frame of 10 minutes. The following figure shows the god view of the environment.

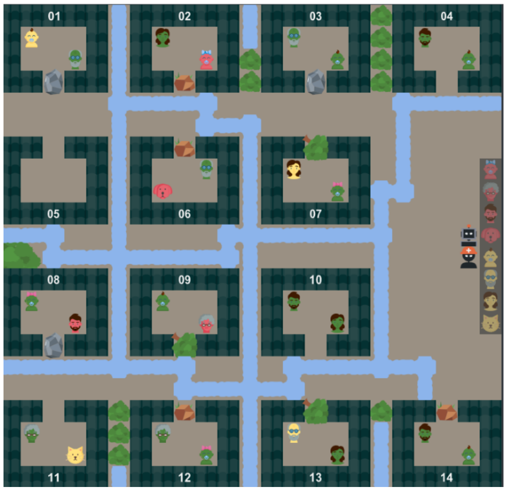


Figure 1: The God view of the environment. It shows different types of victims and obstacles, as well as the drop zone on the right

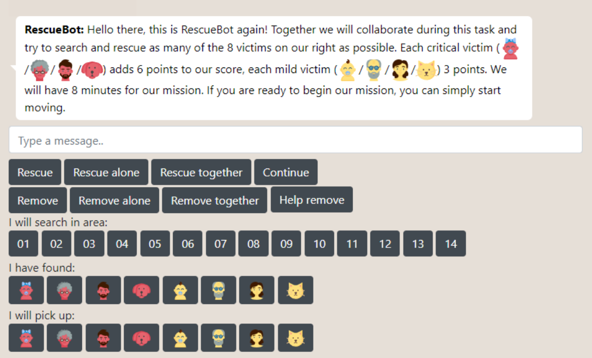
[2] https://matrx-software.com/

As can be seen in Figure 1, victims were placed in various areas, with some areas blocked by one of the three obstacle types; trees, stones, or rocks. In order to enter the areas, the blocking obstacles needed to be removed either alone or together with RescueBot. Furthermore, three distinct types of victims were placed in the environment. Rescuing critically injured victims indicated by the color red yielded 6 points, while rescuing mildly injured victims in yellow gave a reward of 3 points. Green-colored victims entailed healthy victims, and they did not need to be rescued. Additionally, removing an obstacle and picking up a victim alone required four times more time compared to when the task was performed jointly with RescueBot. The following Table 1 illustrates the time differences.

	Alone	Together
Pickup mildly-injured victim	4	1
Pickup critically-injured victim	8	2
Remove stone	4	1
Remove tree	8	2
Remove rock	12	3

Table 1: Object removal and victim pickup time in seconds

The environment also contained a chat functionality for the subjects to communicate with RescueBot. The chat interface could be used to direct RescueBot to perform some tasks as well as to inform Rescuebot about the planned actions to be carried out. For example, the human agent could notify which areas to explore and which victims to pick up next. These tasks were performed through a group of predefined buttons below the chat.



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 functionality, together with a group of predefined buttons for a user to communicate with RescueBot

Ultimately, the environment was set up in such a way that three extreme rains arrived throughout the entire duration of the game. More specifically, they arrived at 2-minute, 4-minute, and 6-minute marks, respectively. The storm could only be avoided if the human agent was in one of the areas (rooms) at the time of the storm. If the human agent failed to seek shelter, the consequences were 10 points reduction in the final score, as well as a deduction in playing time. The human agent was notified about the extreme weathers prior to their arrivals in the form of advice from RescueBot. (Further explanation is in section 3.6) Ultimately, it is important to note that RescueBot was not impacted by the extreme rains in any way and could perform tasks normally under the rain.

3.5 Interdependence Relationships

Two different interdependence relationships were considered in this research: complete independence (baseline), and opportunistic interdependence relationship. In this simulation, different interdependence relationships entailed different availability of actions by the agents. More precisely, in the baseline condition, no joint actions were possible but both agents were capable of performing any task on their own. As for the opportunistic interdependence condition, every task could be performed individually (for both agents) but also jointly.

3.6 Procedure

The user study began by completing a tutorial that was designed to familiarize participants with the available commands. Subsequently, general information about the simulation, including the scoring system and the assigned condition, was provided. After this, the participant was asked to start with the official simulation.

Two minutes after the start of the simulation, RescueBot sent an alert message to the human agent, indicating that a heavy storm is approaching and that the human should seek shelter as soon as possible. About twenty seconds later, the heavy storm arrived, followed by feedback from RescueBot indicating that his advice was correct. The simulation was then paused and the participant was asked to fill in a trust survey measuring the level of human trust in the agent. More on this questionnaire can be found in section 3.7.1.

The game was then resumed, and after another two minutes, another weather forecasting advice was sent from RescueBot. This time, the message indicated that light rain was expected to arrive soon, and advised the human agent to continue searching and rescuing victims instead of seeking shelter. Some moments later, a heavy storm would arrive as opposed to the advice given by RescueBot. Immediately after, the bot would send a feedback message, stating that his advice was incorrect. This was followed by the enforcement of the trust repair strategy of providing explanations and expressing regret. Then, the simulation was paused once again for the participant to complete the trust survey.

Another two minutes after the restart, the last forecasting advice was sent from RescueBot, alerting about the incoming heavy rain. (identical to the first advice) Following the arrival of the extreme rain, feedback regarding the right advice was sent, after which the participant was asked to complete the trust survey.

The game was then resumed and the participant was requested to complete the rest of the simulation. After the completion of the simulation, the participant was presented with a collaboration fluency questionnaire which marked the end of the experiment. More on this fluency questionnaire can be found in 3.7.2

Finally, all the important events during the official run are summarized in the following timeline.

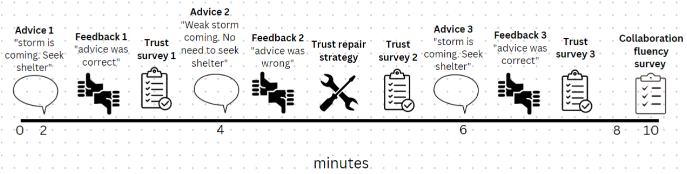


Figure 3: Timeline of the user study

3.7 Measurements

3.7.1 Trust

The level of human trust in RescueBot was measured through a questionnaire whose questions are based on the trust scale for the XAI context proposed by Hoffman [19]. RescueBot can be considered an XAI agent because three weather forecasting suggestions were supported by explanations. Furthermore, the scale measured constructs that are relatively general to automation such as predictability and reliability, suggesting that it was appropriate to use the proposed trust scale in this particular study. The questionnaire consisted of eight questions, all of which use a five-point Likert scale format. That is, an answer to each question was converted to a numeric value in the range from one to five where higher values are assigned to answers indicating a higher level of trust. Each time this questionnaire was filled in by a participant, the average score of the eight questions was computed. The questionnaire was filled out at three different times in a user study. Thus, the computed averages at each of the three timestamps in the same run were compared against each other to understand the impact of trust violations and trust recovery.

3.7.2 Collaboration Fluency

As a subjective fluency metric, collaboration fluency was measured at the end of each user study through a seven-point Likert scale questionnaire. There were eight questions in total, all of which were taken from the paper by Hoffman [5]. The first three measured collaboration fluency directly and the other five questions evaluated downstream outcomes of collaborative fluency. Similar to the trust questionnaire, an answer to each question was converted to a numeric value between one and seven where seven indicates the highest level of fluency perceived by human users. Using the converted numeric values, the average score was computed for each participant. After completing all the user studies, the scores for each participant were aggregated to produce the average scores for the two interdependence conditions.

Some metrics were also logged during the experiments to measure fluency using objective fluency metrics. This includes total task time and the ratio of RescueBot's idle time. Total task time was chosen over score as the score is greatly influenced by the point reduction caused by the rains regardless of the subject's true intention to take shelter.

Robot idle time consists mainly of the time that RescueBot spends waiting for humans to respond to their messages and the time RescueBot spends waiting for humans to come help remove/rescue. Therefore, one can interpret higher robot idle time as an inefficient use of the autonomous agent.

As for total task time, it denotes the amount of time that was required before all the tasks were completed or the time limit was reached. There has been no research done that ties this metric to fluency, however, this was included to serve as a possible justification for the observed fluency value.

4 Responsible Research

4.1 Ethics

In any research, one needs to critically reflect on ethical considerations presented with their study. Regarding this specific study, attention to ethical concerns needs to be paid when conducting the user study. For this reason, we provided each participant with an informed consent form prior to the start of every user study. The following discusses important aspects that were addressed in the consent form.

Firstly, the form informed the purpose of the research and the motivations behind performing the user study. The form emphasized that there are no known risks associated with the study, explaining how our research does not involve collecting personally identifiable data and that the data will be used for non-commercial purposes only. It also highlighted how the anonymity of data is ensured and that the participants can make data removal requests if desired. Lastly, the email address of our supervisor was provided, in case of questions or complaints.

After reading the consent form, participants underwent a series of questions to confirm that they give full consent to how the research is done and how the collected data is addressed. The study only proceeded if the participant agreed to all the presented terms.

4.2 Reproducibility

This project ensures that our experimental setups are easily reproducible. The GitHub repository has been made available for this purpose.[3] It contains five different branches, each corresponding to one of the five interdependence conditions.

Having said that, one can argue that reproducing identical results we obtained may not always be attainable. As will be discussed later in 6.3, different factors such as variations in arrow key press speeds across different participants make the results very susceptible to the selection of participants. Furthermore, some subjects may have a higher level of trust in autonomous agents in general. These factors make it almost impossible to reproduce identical results in the future.

5 Results

5.1 Trust

To analyze the influence of different interdependence relationships (between-subjects-factor) and time (within-subjects-factor) on our dependent variable, namely the trust level, two-way mixed ANOVA was conducted. Datanovia presents five assumptions considered in performing two-way mixed ANOVA [4], out of which the normality assumption, homogeneity of variances assumption, assumption of sphericity,

[3] https://github.com/mawakeb/CSE3000-2023-trust-repair/tree/CSE3000-23

[4] https://www.datanovia.com/en/lessons/mixed-anova-in-r/check-assumptions

and homogeneity of covariances assumption were found to be met from the obtained data. Furthermore, despite the fact that some outliers were detected in the data, there was only one outlier that was marked extreme. This extreme outlier was decided to be removed from my data as including it made a significant impact when analyzing the effect of interdependence relationships on the trust level at different times. The following Figure 4 illustrates the obtained result of the experiment.

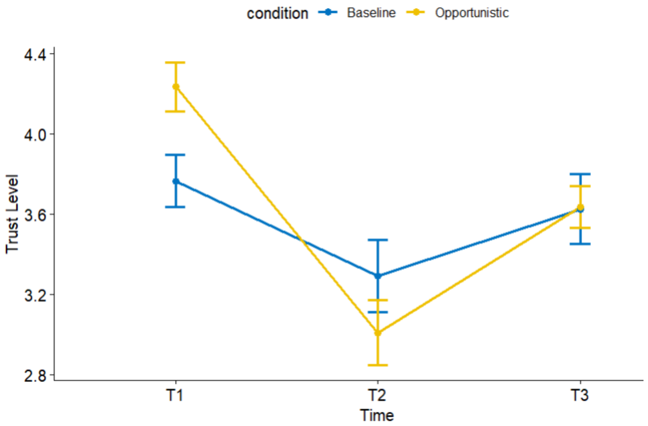


Figure 4: Analyzing the impact of the opportunistic interdependence relationship on the trust level, compared against the baseline result

By analyzing the figure, one can observe a notable decrease in the trust level between T1 and T2 (i.e. before and after trust violation) for the opportunistic interdependence condition compared to the baseline. On top of that, the opportunistic interdependence condition also experienced a greater increase in the trust level between T2 and T3, suggesting effective trust repair during the tasks.

Effect	p-value
Condition	$6.87 * 10^{-1}$
Time	$1.98 * 10^{-7}$
Condition:Time	$1.40 * 10^{-2}$

Table 2: Result of performing two-way mixed ANOVA to analyze the impact of interdependence and time on trust

Additionally, by performing a two-way mixed ANOVA test (Table 2), one can also observe statistically significant two-way interactions between time and conditions on the trust level. This significant two-way interaction can then be decomposed into simple main effects, where in case of significant results, simple pairwise comparisons can be performed.

Firstly, independent samples t-test was performed to compare the means of the trust level of the two interdependence relationships at each time point. Assumptions were checked, confirming the normal distribution of the data and the absence of extreme outliers that, including/excluding them, lead to a significant change in the analysis result. The equality of variances was also checked for each time point and it was found that the variances of the interdependence conditions were not significantly different for all time points, with the p-values being greater than 0.05. This led us to use the standard Student's t-test, which works under the assumption that the variances of the two conditions are equal. Running the test showed that the mean trust level at T1 was significantly different between the conditions with a p-value of 0.0136, but not for T2 and T3 with p-values of 0.2566 and 0.9542 respectively.

Subsequently, the effect size of the interdependence on trust at T1 needed to be analyzed. Cohen's d for Student t-test. was initially considered to be used, however, due to a small sample size of less than 50, Hedge's Corrected version of the Cohen's d was used to prevent us from obtaining an exaggerated effect size. A large effect size was observed, with a d value (effect size) found to be -0.951.

On the other hand, a simple main effect of time on trust was found by conducting one-way repeated measures ANOVA. For the baseline condition, the ANOVA test gave a p-value of 0.031, indicating the presence of statistically significant differences in the trust level at different times. However, conducting a pairwise comparison showed that the adjusted p-values were greater than 0.05 for all possible pairs, suggesting that no significant trust violation or trust repair was observed for the baseline condition. For the opportunistic interdependence condition, it was also found that there is a significant difference in the trust level at different times. ($p = 1.06 * 10^{-7}$) Furthermore, the result of performing pairwise comparisons is made available in Table 3. One can observe a very significant impact of trust violation between T1 and T2 whereas trust recovery between T2 and T3 is significant but slightly less significant compared to the impact of the trust violation.

Group1	Group2	Adjusted p-value
T1	T2	$7.56 * 10^{-5}$
T2	T3	$3.00 * 10^{-3}$

Table 3: Result of performing a pairwise comparison to investigate the simple main effect of time on the trust level for opportunistic interdependence relationship

5.2 Collaboration Fluency

5.2.1 Subjective Fluency Metrics

In this paper, the independent sample t-test was used to determine the impact of the two interdependence relationships on the level of collaboration fluency measured using a questionnaire. The outliers, the normality of data, and the equality of variances were checked, which made sure that there is no violation of the assumptions before performing the analysis.

As can be seen from Figure 5, the mean values were similar between the two conditions (baseline: 5.81, opportunistic: 5.74). Indeed, observing the result of the independent sample t-test analysis gave a p-value significantly greater than 0.05 ($p = 0.7144$), indicating no significant difference in collaboration fluency between the conditions.

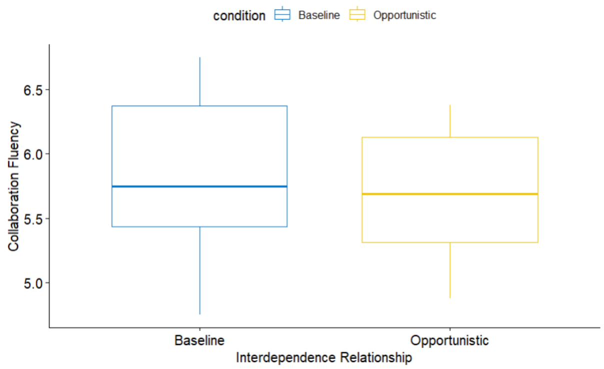


Figure 5: Analyzing the effects of opportunistic interdependence relationship and baseline condition on collaboration fluency

5.2.2 Objective Fluency Metrics

Ratio of RescueBot Idle Time: Figure 6 below shows the differences in RescueBot idle time percentage between the two interdependence conditions. For the baseline condition, the mean value was 0.352 with a standard deviation of 0.0847 whereas for the opportunistic interdependence relationship, the mean value was 0.439 with a standard deviation of 0.0991. Performing an independent sample t-test after checking for assumptions indicated that there is a statistically significant difference between the two idle time percentages. ($p = 0.018$)

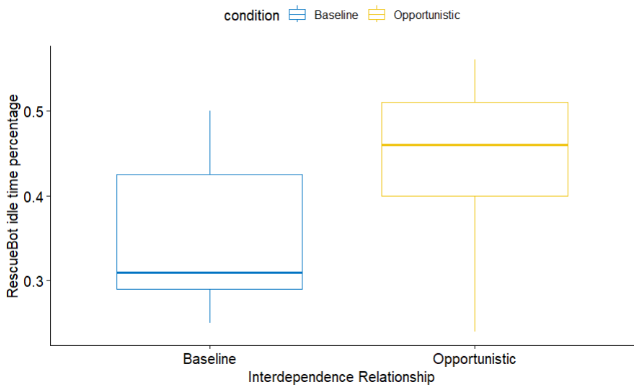


Figure 6: Average Rescuebot's idle time percentage for baseline and opportunistic interdependence relationship

Total Task Time: The mean value for the baseline was 494 seconds with a standard deviation of 65.9 seconds whereas for the opportunistic interdependence, the mean was computed to be equal to 461 seconds, with a standard deviation of 52.3 seconds. (See Figure 7) Subsequently, after confirming that all assumptions were met, an independent sample t-test was performed. The result showed that there is no statistically significant difference between the two conditions ($p = 0.136$)

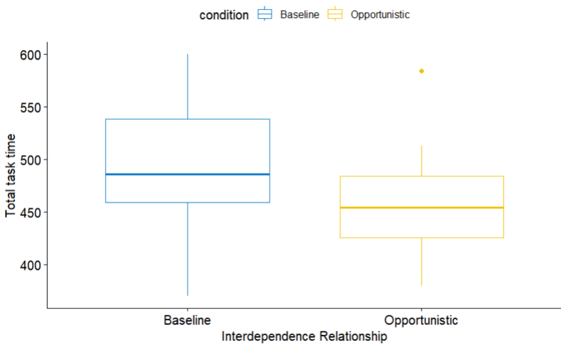


Figure 7: Average total task time for baseline and opportunistic interdependence relationship

6 Discussion

This paper investigated the impact of opportunistic interdependence on trust violation, trust repair, and on collaboration fluency. The following subsections provide a detailed explanation of the results while comparing them against previous work. Finally, the limitations of the experimental setup will be elaborated.

6.1 Trust Violation and Trust Repair

Firstly, investigating the simple main effect of time on the level of trust showed that there is a significant difference in the trust level at T1, with the opportunistic condition having a notably high trust level compared to that of the baseline. This is in line with the research done by Walliser et al. which found that having a team structure increases trust [12]. In our experiment, since no joint action was permitted and each agent was perfectly capable of performing all the tasks on their own, essentially no interdependence existed between the agents in the baseline condition. This implies that the human agent is more likely to consider RescueBot as a separate independent entity rather than their teammate. On the other hand, with soft interdependence, there is an opportunity to perform joint activities with RescueBot, leading to the perception of team structure. Our result, however, contradicts the findings of Verhagen et al. which discovered that a high level of interdependence significantly decreases trust [11].

Furthermore, investigating a simple main effect of time on trust showed that the trust level was not significantly impacted by the trust violation and trust recovery for the baseline condition. As mentioned above, in the baseline, the two agents were under complete independence in terms of their competence to be able to perform the tasks. Therefore, it can be speculated that the participants did not perceive trust in RescueBot as an important factor that influences the performance of the team. Consequently, their trust in RescueBot did not get influenced by the trust violation and the trust repairing process, leading to a relatively consistent trust level throughout the experiment.

In contrast, focusing on the opportunistic interdependence condition, it was found that both the trust violation and trust repair had a significant impact on the level of trust. The significant fluctuation in the trust level can be attributed to the presence of interdependence in the opportunistic condition. In fact, this result is in line with the paper by Johnson and Bradshaw in which they claim that interdependence supports the active and continuous exploration of trust, which ensures the appropriate judgment of the other agent's trustworthiness, leading to maximizing performance [10]. Therefore, because there exists a higher level of interdependence between the agents in the opportunistic condition compared to the baseline, the human agent consistently attempts to calibrate trust with the purpose of making an accurate judgment of whether to collaborate when needed. Finally, one can speculate the reasons behind trust recovery being slightly less significant compared to the impact of trust violation to be attributed to the fact that the human agent is still possible to perform tasks individually in case the trust is lost, without relying on RescueBot. This may give the human agent less incentive to recover lost trust, whereas with a higher level of interdependence relationship, in which some actions strictly require collaboration, there is a bigger motivation to restore trust.

6.2 Collaboration Fluency

Based on subjective fluency metrics, the study has found that the opportunistic condition resulted in a slightly lower level of fluency compared to the baseline but not significantly so. Furthermore, the objective fluency metrics result has shown that, even though there was no significant difference in the total task time, there was a statistically significant difference in the robot idle time percentage between the two conditions. More specifically, the idle time percentage was significantly higher for the opportunistic condition than the baseline. These are somewhat in line with the findings by Hoffman, which stated that robot idle time is consistently but not significantly inversely correlated with subjective fluency perception [5]. However, this non-significant result is not in line with findings from past research, which have found that HATs with higher interdependence level leads to positive outcomes such as improved subjective ratings of team relationship [12]. The reasons behind this can be attributed to the way the experiment was set up, which is explained more in detail in the following paragraphs.

Firstly, the unexpectedly low score for subjective fluency metrics from the opportunistic condition could be explained by the fact that performing joint actions were perceived to be challenging for novice subjects. In fact, it was often observed that the participants struggled to figure out how to seek assistance from RescueBot, especially when they wanted the bot to come and rescue a victim together. One can speculate that this left a negative perception of collaboration in the participants, leading to a lower fluency score.

Secondly, conducting further analysis indicates that collaboration fluency is influenced not only by interdependence relationships but also by whether the incorrect weather forecast caused the participant to be punished by the rain. In fact, while only 60% of the subjects got punished by the rain after the incorrect advice by RescueBot, 86% of the subjects got hit by the rain in the opportunistic interdependence condition. The difference is not surprising given the research which found that interdependence leads to the human being more open to influence from the teammate, and thinking that the information from the teammate is of higher quality [16]. For both the baseline and the opportunistic condition, the average fluency value for those who got punished by the rain after the incorrect weather forecasting advice was lower than for those who managed to take shelter. (Baseline: 5.75 and 5.90, Opportunistic: 5.65 and 6.06 respectively)

Thirdly, and most importantly, the way in which robot idle time was defined did not take into account the fact that idle time arising from the time that RescueBot spends waiting for the human agent to come and help remove/rescue is not applicable to the baseline condition, where no collaboration was possible. This suggests that, for the opportunistic condition, more collaboration led to a higher robot idle time, and therefore, a lower collaboration fluency. In order to avoid this, the idle time should have only included the time it takes for the human agent to respond to the messages, which is equally present in both conditions.

Based on the aforementioned points, one cannot confidently conclude the presence of any significant impact of different interdependence relationships on collaboration fluency. In fact, a detailed analysis with more objective fluency metrics such as human idle time might be required to confidently conclude the relationship between interdependence and collaboration fluency.

6.3 Limitations

There are some other limitations in the way the user study was conducted, which have not been deeply discussed in the previous subsections. This section will explore those considerations more in detail.

Firstly, despite the use of total task time as one of the objective fluency metrics, there were some differences in how the participants controlled the human agent. Some participants spammed arrow keys to move as fast as possible, while others were more gentle with the way they moved. This led to a problem that performance measurements, including lower total task time, are, to some extent, dependent on the speed of controlling the agent.

Secondly, for the opportunistic interdependence condition, the differences in removal/rescue times between performing the tasks individually or jointly were not substantial enough to encourage participants to perform more joint actions. In fact, six out of fifteen participants performed three or fewer joint actions during the whole duration of the game, with some not performing even a single joint action. For those participants, one can speculate that the level of perceived interdependence between the human agent and Resuebot was similar to that of the baseline, making the distinction between the conditions unclear.

Lastly, as was mentioned in the previous subsection, the trust level and collaboration fluency were most likely have been influenced by whether the participants were punished for following the incorrect advice made by RescueBot. In addition, subjects who attempted to but could not take shelter in time and got hit by the rains following the first and third advice might have rated the trust level and the measurement of collaboration fluency lower than those who managed to take shelter. This problem can only be mitigated by classifying participants based on how much they got hit by the rains intentionally or unintentionally, and comparing the results with other participants in the same group.

7 Conclusions and Future Work

The objective of this research was to understand the influence of an opportunistic interdependence relationship on trust violation, trust repair, and on collaboration fluency. A user study was conducted in which participants were requested to complete a search and rescue mission and fill out the trust and fluency questionnaires. Statistical analysis was then performed on the obtained data to gain further insights. This section highlights important findings in the research and discuss recommendations for future research.

Firstly, the opportunistic condition had a significantly higher level of trust before the trust violation, thanks to the perception of team structure in the human agent. In addition, it was found that the opportunistic interdependence condition suffered from a significant loss of trust following the trust violation, but also experienced a relatively significant trust recovery when compared to the baseline condition. This is attributed to the human's perception of higher interdependence, which makes them perform continuous calibration of the teammate's trust level, in order to make correct decisions when they encounter collaboration opportunities. Furthermore, the nature of the soft interdependence condition, in which tasks can still be performed individually without the AI agent, led to a relatively less significant trust recovery compared to the impact of the trust violation.

On the other hand, no statistically significant difference in terms of collaboration fluency was found between the two interdependence conditions based on subjective fluency metrics. Furthermore, for objective fluency metrics, non-significant differences in the results, together with the limitations in the experimental setup which made the metrics unreliable, prevented us from drawing any meaningful conclusion in terms of the relationships between interdependence relationships and collaboration fluency.

As for future work, one could extend this research and investigate if there exists a type of trust repair strategy that performs well against the opportunistic interdependence condition in particular. For this research, we have utilized an identical trust repair strategy for both interdependence conditions. However, HATs could benefit from incorporating a specific type of trust repair strategy depending on the dynamics of the team.

References

[1] W. Bluethmann, R. Ambrose, M. Diftler, S. Askew, E. Huber, M. Goza, F. Rehnmark, C. Lovchik, and D. Magruder, "Robonaut: A robot designed to work with humans in space," *Autonomous robots*, vol. 14, pp. 179–97, 03 2003.

[2] N. McNeese, M. Demir, E. Chiou, and N. Cooke, "Trust and team performance in human-autonomy teaming," *International Journal of Electronic Commerce*, vol. 25, pp. 51–72, 01 2021.

[3] E. Salas, D. Sims, and S. Burke, "Is there a "big five" in teamwork?," *Small Group Research*, vol. 36, pp. 555 –599, 10 2005.

[4] E. S. Kox, J. H. Kerstholt, T. F. Hueting, and P. W. de Vries, "Trust repair in human-agent teams: the effectiveness of explanations and expressing regret," Master's thesis, University of Twente, 2021.

[5] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 209–218, 2019.

[6] G. Hoffman and C. Breazeal, "Cost-based anticipatory action selection for human–robot fluency," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 952–961, 2007.

[7] E. Kox, L. Siegling, and J. Kerstholt, "Trust development in military and civilian human–agent teams: The effect of social-cognitive recovery strategies," *International journal of social robotics*, vol. 14, pp. 1323–1338, July 2022. Funding Information: This material is based upon work supported by the Dutch Ministry of Defense's exploratory research program. Publisher Copyright: © 2022, The Author(s).

[8] D. Ferrin, P. Kim, C. Cooper, and K. Dirks, "Silence speaks volumes: The effectiveness of reticence in comparison to apology and denial for responding to integrity- and competence-based trust violations," *The Journal of applied psychology*, vol. 92, pp. 893–908, 08 2007.

[9] P. H. Kim, K. T. Dirks, C. D. Cooper, and D. L. Ferrin, "When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation," *Organizational Behavior and Human Decision Processes*, vol. 99, no. 1, pp. 49–65, 2006.

[10] M. Johnson and J. Bradshaw, "The role of interdependence in trust," 03 2021.

[11] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman, "The influence of interdependence and a transparent or explainable communication style on human-robot teamwork," *Frontiers in Robotics and AI*, vol. 9, 2022.

[12] J. Walliser, P. Mead, and T. Shaw, "The perception of teamwork with an autonomous agent enhances affect and performance outcomes," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, pp. 231–235, 09 2017.

[13] M. Johnson, J. Bradshaw, P. J. Feltovich, C. Jonker, M. Riemsdijk, and M. Sierhuis, "Coactive design: Designing support for interdependence in joint activity," *Human Robot-Interaction*, vol. 3(1), 03 2014.

[14] J. Lee and K. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, pp. 50–80, 02 2004.

[15] A. Baker, E. Phillips, D. Ullman, and J. Keebler, "Toward an understanding of trust repair in human-robot interaction: Current research and future directions," *ACM*

Transactions on Interactive Intelligent Systems, vol. 8, pp. 1–30, 11 2018.

[16] C. Nass, B. Fogg, and Y. Moon, "Can computers be teammates?," *International Journal of Human-Computer Studies*, vol. 45, no. 6, pp. 669–678, 1996.

[17] M. Johnson, J. Bradshaw, P. J. Feltovich, C. Jonker, B. Riemsdijk, and M. Sierhuis, "Autonomy and interdependence in human-agent-robot teams," *IEEE Expert / IEEE Intelligent Systems - EXPERT*, vol. 27, pp. 43–51, 03 2012.

[18] M. Johnson, J. Bradshaw, P. J. Feltovich, C. Jonker, B. Riemsdijk, and M. Sierhuis, "The fundamental principle of coactive design: Interdependence must shape autonomy," vol. 6541, pp. 172–191, 01 2010.

[19] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance," *Frontiers in Computer Science*, vol. 5, 2023.