HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240203

The Influence of Interdependence on Trust Calibration in Human-Machine Teams

Ruben S. VERHAGEN [a,1], Alexandra MARCU [a], Mark A. NEERINCX [a,b], and Myrthe L. TIELMAN [a]

[a] *Delft University of Technology, Delft, The Netherlands*
[b] *TNO, Soesterberg, The Netherlands*

Abstract. In human-machine teams, the strengths and weaknesses of both team members result in dependencies, opportunities, and requirements to collaborate. Managing these interdependence relationships is crucial for teamwork, as it is argued that they facilitate accurate trust calibration. Unfortunately, empirical research on the influence of interdependence on trust calibration during human-machine teamwork is lacking. Therefore, we conducted an experiment (n=80) to study the effect of interdependence relationships (complete independence, complementary independence, optional interdependence, required interdependence) on human-machine trust calibration. Participants collaborated with a virtual agent during a simulated search and rescue task in teams characterized by one of the four interdependencies. A machine-induced trust violation was included in the task to facilitate dynamic trust calibration. Results show that the interdependence relationships during human-machine teamwork influence perceived trust calibration over time. Only in the teams with joint actions (optional and required interdependence) does perceived trust in the machine not recover to its initial pre-violated value. However, results show that the correlation between perceived trust in the machine and machine trustworthiness is strongest in these teams with joint actions, suggesting a more accurate trust calibration process. Overall, our findings provide some first evidence that interdependence relationships during human-machine teamwork influence human-machine trust calibration.

Keywords. interdependence, trust calibration, human-machine teamwork

1. Introduction

Humans and intelligent machines increasingly work together as teammates on complex tasks such as manufacturing and firefighting [1]. Machines often outperform humans concerning rapid, rational, and repetitive decision-making, whereas humans are usually better at handling uncertainty and unexpected situations [2]. These separate strengths and weaknesses of humans and machines result in different dependencies, opportunities, and requirements to collaborate [3]. The ultimate goal of human-machine teams is to harness the combination of strengths of both humans and machines to accomplish what neither can do alone [4].

[1] Corresponding Author: Ruben Verhagen, r.s.verhagen@tudelft.nl

Several factors determine the success of human-machine teams, for example, effectively managing the interdependence relationships between both team members [5]. Another crucial determinant is appropriate human trust in machines, meaning that they know both the potentials and limitations of machines [6,7]. A lack of appropriate trust (i.e., over- or under-trust) is one of the main reasons for the disuse and misuse of machines. This lack can be corrected by a trust calibration process over time and repeated interactions, allowing humans to adjust their expectations of the machine's reliability and trustworthiness [6,7,8]. During the trust calibration process, repairing trust violations caused by machine errors is more difficult than building trust initially [9,10].

It is argued that interdependence relationships between humans and machines facilitate the assessment of trustworthiness of intelligent machines and accurate trust calibration by humans [11]. However, there is a lack of empirical research on the exact influence of interdependence on trust calibration in human-machine teams. For example, how different interdependence relationships during human-machine teamwork influence the trust calibration process over time is unknown. Therefore, this study investigates how complete independence, complementary independence, optional interdependence, and required interdependence influence human-machine trust calibration. To do this, we conducted a user study where participants collaborated with a virtual agent during a simulated search and rescue task in teams characterized by one of the four interdependencies.

2. Background

2.1. Interdependence in Human-Machine Teams

Interdependence relationships are the complementary relationships humans and machines rely on to manage dependencies during joint activities [3,12]. Joint activities concern situations in which the actions of humans depend on those of machines (and vice versa) over a sustained sequence of actions and towards a shared goal [3]. These joint activities are characterized by required, optional, complementary, or no dependencies between humans and machines, caused by their capabilities to execute actions individually and assist each other during action execution [3].

When humans and machines can both execute actions independently while collaborating towards a shared goal, they are hardly dependent on each other. On the other hand, complementary dependencies between humans and machines exist when each can only execute their unique actions that contribute to completing the overall task. Optional dependencies stem from recognizing opportunities to be more efficient when executing actions jointly rather than independently [2,3]. Finally, required dependencies originate from both team members' lack of knowledge, skills, abilities, and resources to competently execute an action independently, but the potential to assist each other to execute the action jointly [2,3]. This distinction between complete independence, complementary independence, optional interdependence, and required interdependence essentially forms a hierarchy in coordination, dependencies, and strength of the interdependence relationship [2,3]. As these different interdependence relationships heavily affect mutual reliance and dependencies, they play a critical role in the trust relationship between humans and machines [11].

2.2. Trust in Human-Machine Teams

An early definition of trust is believing that someone or something else will act in your best interest and accepting vulnerability to this person's or entity's actions [13]. So, there is a trusting party (the trustor) and a party to be trusted (the trustee) [13]. Here, trust can be considered as the trustor's perception of the trustee's trustworthiness [7,14]. Trust is critical in all circumstances where people are in any way dependent on other's actions, and thus more relevant in high-risk situations [7,13]. More specifically, more trust is required when the perceived risk of relying on someone or something else is higher [13]. We believe that interdependence influences the perceived risk associated with relying on someone or something else and, thus, indirectly, how much trust is required during the relationship. For example, relying on someone who can execute actions you can not is less risky than relying on someone to execute actions jointly.

Instead of blindly trusting machines, human-machine trust must be appropriate [7]. Human-machine trust is appropriate when the human's trust in the machine is equal to the machine's actual trustworthiness [7,15]. This match between trust and trustworthiness involves both trusting trustworthy machines and distrusting untrustworthy machines. When appropriate trust is directly caused by information about the actual trustworthiness of the machine, this is called warranted appropriate trust [16,17]. Fostering appropriate trust is crucial as a lack of appropriate human-machine trust can cause over- or under-trust in and over- or under-reliance on machines, potentially resulting in detrimental outcomes [6,7,18,19]. Fostering appropriate trust involves a process of trust calibration that corrects for over- and under-trust over time and repeated interactions, allowing humans to adjust their expectations of the machine's reliability and trustworthiness [6,7,8].

During the trust calibration process, human-machine trust is rarely stable but instead changes over time based on past and current interactions [27,28,29,30]. Decreases in human-machine trust resulting from machine-induced trust violations can have lasting effects and are hard to recover from [29,30]. To this end, machines can deploy several trust repair strategies to repair human trust after they damage or violate it [9,30,31,32]. The most commonly used trust repair strategies include apologies, denials, explanations, and promises [29,33,34]. The impact of these trust repair strategies on human trust has been mixed, with studies showing positive, no, or even negative effects [34,35]. Moderating factors might explain these mixed results, such as the timing of the repair strategy, violation type, and violation severity [9,34]. One general result, however, seems to be the effectiveness of machine apologies for restoring trust [9,36,37]. Adding an explanation to the apology can even amplify this effect [9,38].

Explanations are not merely a trust repair strategy but also one of the primary methods for fostering appropriate human-machine trust. They specifically aim to make intelligent machines more transparent and understandable to humans [7,20,21]. Examples include machine explanations, confidence scores, and uncertainty communication, providing information about the capabilities and limitations of machines and how and why they make decisions [22,23,24]. Prior literature has shown that these forms of machine transparency can improve appropriate trust in machines [22,23,24,25,26].

2.3. Interdependence for Trust Calibration in Human-Machine Teams

In addition to machine explanations, it is argued that interdependence relationships also play a critical role in the trust calibration process [11]. In order to do so, interdependence relationships need to be supported by observable, predictable, and directable machines [3,11]. This means that intelligent machines should be transparent and understandable enough for humans to reasonably rely on them while also allowing humans to influence their behavior [3,11]. This way, interdependence relationships can support the active and continuous exploration of trust between humans and machines to ensure that human assessments are appropriate for achieving the best possible outcomes [11].

As both trust and interdependence relationships involve risk, reliance, and dependencies, it is unsurprising that interdependence and trust are related [12]. Johnson and Bradshaw [11] argue that interdependence relationships facilitate the assessment of the trustworthiness of the machine and accurate trust calibration required for developing warranted appropriate trust. However, interdependence relationships between humans and machines can vary in terms of coordination and dependencies, such as required or optional dependencies during joint activities [2,3]. So far, there is a lack of empirical research on how these different interdependence relationships during human-machine teamwork influence human-machine trust calibration. Our study will fill that gap by comparing how complete independence, complementary independence, optional interdependence, and required interdependence influence human-machine trust calibration.

3. Method

3.1. Design

We conducted an experiment to investigate the influence of interdependence relationships during human-machine teamwork on human-machine trust calibration. To ensure a dynamic trust calibration process, we added a trust violation caused by incorrect machine advice. The experiment had a 3x4 mixed design with time as the within-subjects independent variable and interdependence as the between-subjects independent variable. Time consisted of three conditions (pre-violation, post-violation, post-recovery) and interdependence of four conditions (complete independence, complementary independence, optional interdependence, and required interdependence). As dependent variables, we measured perceived trust and the appropriate reliance rate at each of the three time points.

3.2. Participants

We recruited 80 participants through personal contacts within the university (29 female and 51 male participants). Sixty-nine participants had an age range of 18-24 years old, seven participants of 25-34 years old, one participant of 35-44 years old, two participants of 45-54 years old, and one participant of 55-64 years old. In terms of education, two participants went to high school but did not obtain a diploma, 44 participants were high school graduates, nine participants obtained some college credit but no degree (yet), one participant obtained an Associate degree, 19 participants obtained a Bachelor's degree, and five participants obtained a Master's degree. Concerning gaming experience, 11 participants had no experience at all, 19 participants had a little, nine participants had a moderate amount, 22 participants had a considerable amount, and 19 participants had a lot. All participants signed an informed consent form before participating in the study, approved by the ethics committee of our institution (ID 3002). Since each participant was assigned to one of the four interdependence conditions, it was essential to control for gender, age, education, and gaming experience between these conditions. Results showed no significant differences between interdependence conditions for any of the demographic factors gender (χ^2 (3) $= 3.62, p = 0.31$), age ($W = 1.23, p = 0.75$), education ($W = 3.94, p = 0.27$), and gaming experience ($W = 0.86, p = 0.84$). Therefore, we did not further control for these demographics during data analysis.

3.3. Hardware and Software

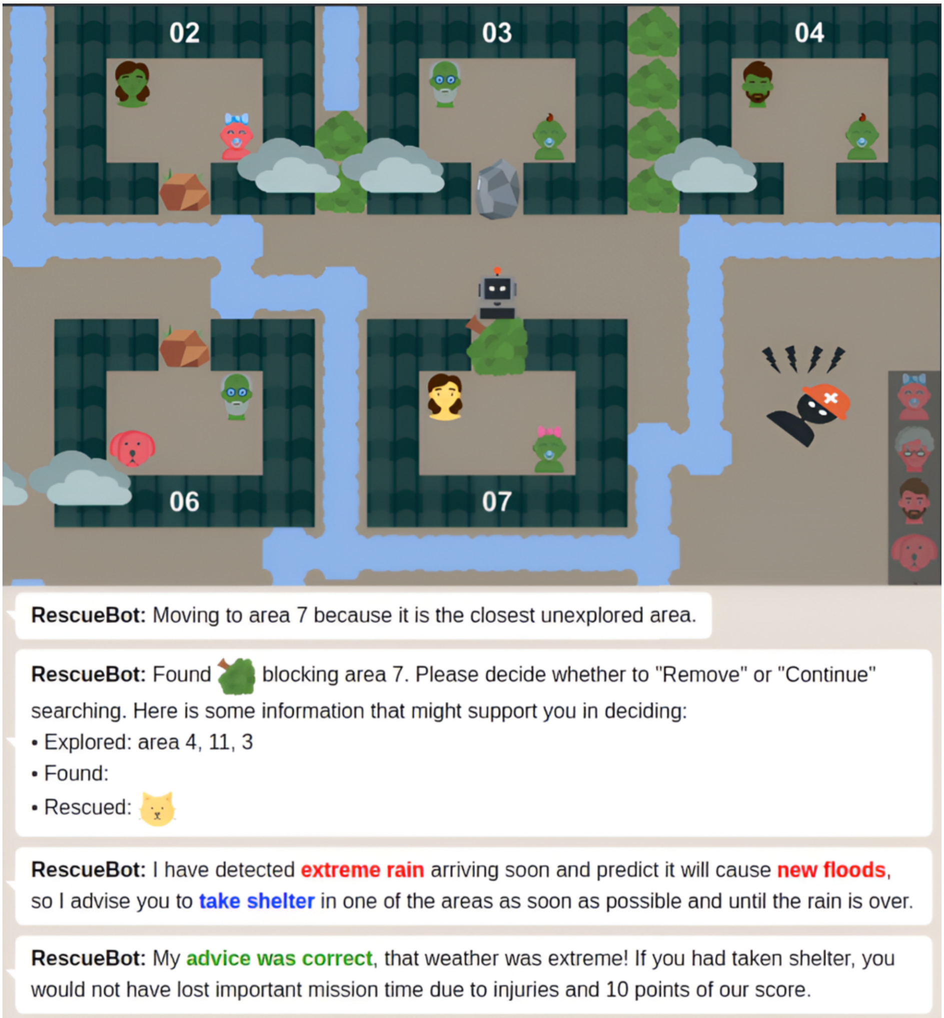
To run this experiment, we used a laptop and the Human-Agent Teaming Rapid Experimentation (MATRX) software, a Python package for facilitating human-agent teaming research (https://matrx-software.com/). The laptop was used to launch our two-dimensional grid world created using MATRX. All subjective measures were collected using Qualtrics, while all objective measures were automatically logged using MATRX.

3.4. Environment

We built a MATRX world consisting of 14 areas, 26 collectable objects, 12 obstacles, and one drop zone (see Figure 1 for part of the world). Furthermore, we added an autonomous virtual agent (RescueBot) and a human agent (controlled by the participants) to our world. We designed an environment in which these two agents had to collaborate during a search and rescue task. To ensure an inclusive and realistic victim representation, we created the following eight victim types making up the world's collection goal: girl, boy, woman, man, older woman, older man, cat, and dog. In addition, we created three injury types: critical, mild, and healthy. Injury type was represented by the color of the victims, where red reflected critically injured, yellow mildly injured, and green healthy victims. Eight of the 26 victims were either mildly or critically injured and had to be delivered at the drop zone, whereas the other 18 were healthy. We also added three obstacle types in front of area entrances: boulder, tree, and stone. Finally, we added flooded water to the environment, which slowed the agents' speed as they moved through it.

3.5. Task

The objective of the task was to find the target victims in the different areas and carry them to the drop zone. Interdependence relationships between humans and RescueBot were manipulated, resulting in four conditions characterized by unique dependencies [3]. In the complete independence condition, the human and RescueBot could execute all actions independently (i.e., remove all obstacles and rescue all victims). In the complementary independence condition, RescueBot could only remove obstacles, whereas the human could only rescue victims. The other two conditions also included joint actions. In the optional interdependence condition, the human and RescueBot could execute all actions independently and jointly. However, joint action execution was four times faster than independent action execution. In the required interdependence condition, all actions had to be executed jointly. Independently removing obstacles took four seconds for stones, eight seconds for trees, and 12 seconds for boulders. Independently rescuing victims took four seconds for mildly injured victims and eight seconds for critically injured victims. Participants had ten minutes to complete the task (i.e., drop all victims at the drop zone) and received points for each victim they rescued. Rescuing critically injured



RescueBot: Moving to area 7 because it is the closest unexplored area.

RescueBot: Found 🌳 blocking area 7. Please decide whether to "Remove" or "Continue" searching. Here is some information that might support you in deciding:
- Explored: area 4, 11, 3
- Found:
- Rescued: 🐱

RescueBot: 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.

RescueBot: My **advice was correct**, that weather was extreme! If you had taken shelter, you would not have lost important mission time due to injuries and 10 points of our score.

Figure 1. Experimenter view of the MATRX world used for our study.

victims added six points to the total score, while rescuing mildly injured victims added three points, resulting in a maximum possible score of 36 points. Other than points and rescue time, no other differences existed between mildly and critically injured victims.

During the task, extreme rain hit the MATRX world three times: after two, four, and six minutes. This rain lasted for ten seconds and if participants did not seek shelter in one of the areas during the rain, they would lose ten points of their score and their avatar would freeze until the rain disappeared. The extreme rain merely affected score and time; it did not affect the victims to be rescued. Before the extreme rain, RescueBot warned the participants about its severity and correspondingly recommended seeking shelter or continuing with the search and rescue task. Each message was accompanied by a ping sound and color highlights to draw attention. After the rain disappeared, RescueBot provided feedback on whether the advice was correct, and more flooded water was added to the environment. RescueBot's first advice was correct. In contrast, RescueBot's second

Table 1. Overview of the advice and feedback messages provided by RescueBot during the experiment.

Message type	Message content
Advice T1,3	I have detected extreme rain arriving soon and predict it will cause new floods.
	I advise you to take shelter in one of the areas as soon as possible, until the rain is over.
Feedback T1,3	My advice was correct, that weather was extreme! If you had (not) taken shelter,
	you would (not) have lost 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.
	I advise you to continue searching and rescuing victims.
Feedback T2	My advice was wrong. The amount of rain was heavy instead of light.
	Because of that my flood prediction was incorrect. I am really sorry.

advice was incorrect, provoking a trust violation. Therefore, the following feedback message contained a trust repair message explaining what happened and expressing regret [9]. We included this element of risk to the task because risk and vulnerability are critical elements of trust [7]. RescueBot's third recommendation was correct again. Table 1 shows all the advice and feedback messages provided by RescueBot.

3.6. Agent Types

We added two agents to the world: an autonomous rule-based virtual agent (RescueBot) and a human agent controlled by the participants using their keyboards. RescueBot always moved to the closest unsearched area during the search and rescue task. Furthermore, it tracked which areas the team had searched, which victims the team had found and where, and which victims the team had rescued. RescueBot did not execute any removing or rescuing actions autonomously. Instead, it asked the participants to decide whether to remove obstacles or rescue victims independently or jointly, accompanied by a summary of the explored areas, found victims, and rescued victims (see Figure 1). This way, RescueBot's behavior was consistent for all interdependence conditions.

Both agents could only carry one victim at a time (either independently or jointly), detect each other within two grid cells, detect and remove obstacles or pick up victims within one grid cell, and detect walls and doors from anywhere. Both agents could also communicate using the chat box shown in Figure 1. Using buttons, participants could share their actions, perceptions, assistance requests, and answers to any questions asked by RescueBot. RescueBot added the shared information to its memory and adjusted its behavior correspondingly (e.g., by not moving to the same areas as the participants).

3.7. Measures

We used self-reporting and behavior to measure perceived trust in and demonstrated reliance on RescueBot [7]. More specifically, we subjectively measured perceived user trust in RescueBot using the 5-point Likert scale for trust in explainable artificial intelligence systems [39]. This scale consisted of eight items and measured confidence in and predictability, reliability, safety, efficiency, wariness, performance, and likeability of RescueBot. We calculated the mean of these eight items as the final perceived trust score for each of the three time points separately.

In addition, we objectively logged whether participants followed the advice given by RescueBot. Based on this data, we calculated the appropriate reliance rate on RescueBot. *Appropriate reliance* was defined as appropriate reliance on RescueBot's correct advice at T1 and T3 and appropriate non-reliance on RescueBot's incorrect advice at T2. Accordingly, we calculated the appropriate reliance rate at each time point by dividing the number of appropriate (non-)reliance occurrences by the number of received recommendations so far. This way, the appropriate reliance rate was a cumulative variable.

3.8. Procedure

Participants first completed a tutorial to familiarize them with the environment, controls, and messaging system. Next, participants started the actual experiment. After one minute and 45 seconds, RescueBot warned the participants about arriving rain and whether to seek shelter. After two minutes, the rain arrived and lasted for ten seconds. When the rain disappeared, RescueBot provided feedback on whether its advice was correct. After two minutes and 20 seconds, the game paused, and participants were asked to fill out the trust questionnaire for the first time. This cycle of warning, rain, feedback, and trust questionnaire was repeated two more times with similar intervals, with the other warnings arriving at three minutes and 45 seconds and five minutes and 45 seconds. The whole study lasted about 30 minutes and was conducted offline.

4. Results

4.1. Perceived Trust and Appropriate Reliance

To investigate the effects of interdependence and time on perceived trust in RescueBot (Figure 2A), we conducted both a parametric and nonparametric mixed ANOVA. We conducted both ANOVAs because the assumption of homogeneity of variances for the parametric mixed ANOVA was slightly violated at T3. Results of the parametric mixed ANOVA showed a statistically significant interaction between interdependence and time on perceived trust ($F(6, 152) = 2.83, p < 0.025, \eta^2_G = 0.042$). Results showed that the simple main effect of interdependence on perceived trust was not significant at any of the time points. In contrast, results showed that the simple main effect of time on perceived trust was significant for complete independence ($F(2, 38) = 11.1, p < 0.001, \eta^2_G = 0.18$), complementary independence ($F(2, 38) = 9.45, p < 0.005, \eta^2_G = 0.16$), optional interdependence ($F(1.38, 26.2) = 35.6, p < 0.001, \eta^2_G = 0.37$), and required interdependence ($F(1.27, 24.2) = 35.4, p < 0.001, \eta^2_G = 0.50$). Pairwise t-test comparisons using a Bonferroni correction revealed significant differences in trust scores between all time points and for all interdependencies, except between T1 and T3 for complete independence and complementary independence (Table 2 and Table 3).

To confirm these results, we ran the nonparametric rank-based mixed ANOVA [40]. Again, results showed a statistically significant interaction between interdependence and time on perceived trust ($F(4.56) = 2.29, p < 0.05$, effect size $= 0.44$). These results also showed that the simple main effect of interdependence was not significant at any of the time points. Moreover, the results again showed that the simple main effect of time on perceived trust was significant for complete independence ($\chi^2(2) = 13.40, p < 0.0025, W = 0.36$), complementary independence ($\chi^2(2) = 14.50, p < 0.001, W = 0.34$), optional interdependence ($\chi^2(2) = 30.30, p < 0.001, W = 0.76$), and required interdepen-

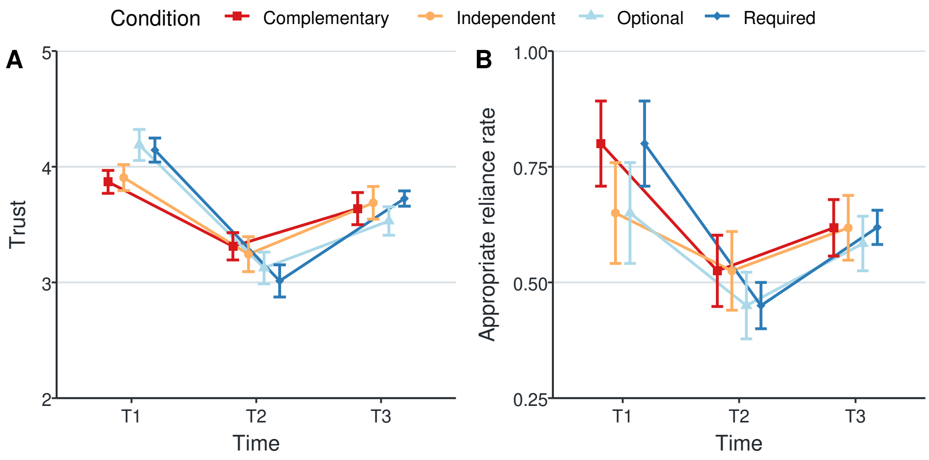


Figure 2. Interaction plots of the effects of interdependence and time on perceived trust (A) and the appropriate reliance rate (B). Error bars represent the standard errors.

Table 2. Pairwise t-test and Wilcoxon comparisons for the simple main effect of time on perceived trust for each interdependence condition. Bold values show the non-significant pairwise comparisons.

Condition	Time points	Δ mean	*t*	*p*	*W*	*p*
Complete independence	T1 vs. T2	-0.67	3.98	< 0.005	129	< 0.01
	T1 vs. T3	-0.22	1.66	**0.34**	109	**0.39**
	T2 vs. T3	+0.45	-3.46	< 0.001	21	< 0.01
Complementary independence	T1 vs. T2	-0.56	4.02	< 0.005	172	< 0.01
	T1 vs. T3	-0.23	1.59	**0.38**	146	**0.39**
	T2 vs. T3	+0.33	-3.37	< 0.025	23	< 0.05
Optional interdependence	T1 vs. T2	-1.06	6.50	< 0.001	207	< 0.001
	T1 vs. T3	-0.66	7.10	< 0.001	208	< 0.001
	T2 vs. T3	+0.40	-3.53	< 0.01	16	< 0.01
Required interdependence	T1 vs. T2	-1.13	6.35	< 0.001	210	< 0.001
	T1 vs. T3	-0.41	4.71	< 0.001	150	< 0.01
	T2 vs. T3	+0.72	-5.65	< 0.001	0	< 0.001

dence ($\chi^2(2) = 35.7, p < 0.001, W = 0.89$). Finally, pairwise Wilcoxon comparisons using a Bonferroni correction also revealed significant differences in trust scores between all time points and for all interdependencies, except between T1 and T3 for complete independence and complementary independence (Table 2 and Table 3).

To investigate the effects of interdependence and time on the appropriate reliance rate (Figure 2B), we conducted the nonparametric mixed ANOVA because of not normally distributed data. Results showed a significant main effect of time on the appropriate reliance rate ($F(1.35) = 48.06$, $p < 0.001$, effect size $= 1.10$). Pairwise Wilcoxon comparisons using a Bonferroni correction revealed significant differences between the appropriate reliance rates at T1 and T2 ($p < 0.001$) and T2 and T3 ($p < 0.001$).

Table 3. Descriptive statistics for each combination of time and interdependence condition. M refers to the mean, MR to the mean rank, SD to the standard deviation, and AR% to the appropriate reliance rate.

Condition	Time	M (SD) trust	MR (SD) trust	M (SD) AR%	MR (SD) AR%
Complete	T1	3.91 (0.50)	153.43 (61.72)	0.65 (0.49)	136.08 (85.88)
independence	T2	3.24 (0.68)	82.23 (66.69)	0.53 (0.38)	103.68 (68.37)
	T3	3.69 (0.63)	128.00 (70.54)	0.62 (0.31)	121.80 (60.06)
Complementary	T1	3.87 (0.44)	147.40 (55.40)	0.80 (0.41)	162.40 (72.02)
independence	T2	3.31 (0.53)	85.05 (51.26)	0.53 (0.34)	101.35 (62.34)
	T3	3.64 (0.62)	122.95 (66.19)	0.62 (0.27)	120.10 (55.08)
Optional	T1	4.19 (0.60)	178.25 (64.27)	0.65 (0.49)	136.08 (85.88)
interdependence	T2	3.13 (0.61)	70.58 (55.53)	0.45 (0.32)	87.03 (55.14)
	T3	3.53 (0.55)	110.93 (59.77)	0.58 (0.26)	112.88 (53.85)
Required	T1	4.14 (0.47)	177.78 (54.32)	0.80 (0.41)	162.40 (72.02)
interdependence	T2	3.01 (0.62)	60.60 (58.84)	0.45 (0.22)	82.38 (35.88)
	T3	3.73 (0.30)	128.83 (39.17)	0.62 (0.17)	119.85 (37.21)

4.2. *Effects of Interdependence on Reliance and Injuries*

Next, we investigated if the interaction between interdependence and time on perceived trust (Figure 2A) could be explained by differences between interdependence conditions in the number of injuries or how much they relied on RescueBot. Here, the underlying assumptions were that more reliance could result in more trust [7], and more injuries (and thus lost points) in less trust. However, the already reported nonparametric mixed ANOVA only showed a significant main effect of time on the appropriate reliance rate. Results of another nonparametric mixed ANOVA also showed a non-significant interaction effect of interdependence and time on the general reliance rate ($F(3.95) = 0.83$, $p = 0.51$, effect size $= 0.26$), and non-significant main effect of interdependence on the general reliance rate ($F(2.96) = 1.77$, $p = 0.15$, effect size $= 0.26$). Finally, results showed that all interdependence conditions were homogeneous concerning how often they were injured by the rain (χ^2 (3) $= 0.21$, $p = 0.98$), also at T1 (χ^2 (3) $= 2.26$, $p = 0.52$), T2 (χ^2 (3) $= 4.80$, $p = 0.19$), and T3 separately (χ^2 (3) $= 2.35$, $p = 0.50$).

4.3. *Accuracy of the Trust Calibration Process*

Finally, for each interdependence condition, we compared the trust calibration process over time with RescueBot's actual trustworthiness over time, expressed in terms of its advice accuracy [7,41]. More specifically, RescueBot's advice accuracy was 100% at T1, 50% at T2, and 67% at T3. For each interdependence condition, we ran a Spearman's rank-order correlation to assess the relationship between perceived trust in RescueBot and advice accuracy of RescueBot. Results showed a statistically significant positive correlation between perceived trust and advice accuracy for complete independence ($\rho = 0.42$, $p < 0.001$), complementary independence ($\rho = 0.40$, $p < 0.005$), optional interdependence ($\rho = 0.60$, $p < 0.001$), and required interdependence ($\rho = 0.69$, $p < 0.001$).

5. Discussion and Conclusion

5.1. Discussion

Our results show that interdependence relationships during human-machine teamwork influence human-machine trust calibration over time (Figure 2A). Across all interdependence relationships, we observe significant post-violation trust decreases compared to pre-violated trust (T2 vs. T1) and significant post-recovery trust repairs compared to post-violated trust (T3 vs. T2). However, only in the teams with joint actions (optional and required interdependence) we observe a significant post-recovery trust decrease compared to pre-violated trust (T3 vs. T1). In other words, human-machine trust does not recover to its initial pre-violated value only in the teams with joint actions (Section 4.1). Since we do not find evidence for an influence of interdependence on reliance or the number of injuries (Section 4.2), this finding can more likely be attributed to the direct influence of interdependence relationships on human-machine trust calibration.

The results further indicate that the correlation between perceived trust in RescueBot and RescueBot's advice accuracy is significant for all interdependence relationships but strongest for the teams with joint actions (Section 4.3). This finding supports Johnson and Bradshaw's claim [11] that interdependence facilitates accurate trust calibration. However, it also extends the claim by showing that stronger interdependence relationships with joint actions facilitate more accurate trust calibration aligning with RescueBot's trustworthiness. This might explain why human-machine trust does not recover to its initial pre-violated value in the teams with joint actions.

We believe that the perceived risk associated with relying on machines [13] increases with the strength of the interdependence relationship, and therefore, more trust is necessary for human-machine teams with joint actions. Prior research has shown that under such conditions of increased trust necessity, over-trust can be promising for trust calibration [7,42]. Therefore, we speculate that over-reliance on the incorrect advice at T2 resulted in a more accurate trust calibration in the teams with higher trust necessity caused by joint actions. This might also explain why the stronger interdependence relationships with joint actions facilitate more accurate trust calibration aligning with RescueBot's trustworthiness. However, follow-up research is required to support these hypothesized relationships between interdependence, risk, (over-)reliance, and trust (necessity).

Finally, we did not find evidence of an effect of interdependence on the calibration of appropriate human-machine reliance. However, timing was an important distinction between perceived trust and the appropriate reliance rate, as perceived trust was recorded after the consequences of reliance behavior. Therefore, it made little sense to compare the calibration of appropriate reliance with RescueBot's actual trustworthiness over time, as participants could not make an informed estimate of its accuracy at T1. All in all, our results highlight that interdependence relationships are crucial to consider carefully in human-machine teams as they can influence perceived human-machine trust calibration.

5.2. Limitations and Future Work

We identify a few limitations of our study. First, we only used three time points to reflect human-machine trust calibration over time, which is a simplified representation. Even though this representation aided in capturing some critical aspects of the calibration process, the limited temporal scope probably did not capture all nuanced aspects of trust calibration over time. Therefore, future research could increase the temporal scope of the study, facilitating a more detailed investigation of the trust calibration process.

Furthermore, we used four distinctive interdependence relationships for our interdependence conditions. Again, this is a simplified representation of human-machine collaboration, which is often characterized by a mix of all four relationships [3,43]. However, using these four distinctive relationships allowed us to examine their unique influence on trust calibration. Even though human-machine teamwork often involves a mix of all interdependencies, our results still provide developers with crucial insights. For example, how violated trust does not recover to its initial value for teams engaged in joint actions and that these teams demonstrate a more accurate trust calibration.

We identify several directions for future work. For example, investigating the interaction between interdependence and trust repair strategy on trust calibration. We speculate that specific repair strategies work better for certain interdependencies, such as promises for relationships with joint actions and explanations for independent collaboration. Future work could test these hypotheses by extending our research environment with different trust repair strategies [29,33,34]. These results could provide valuable insights allowing machines to adapt their trust repair strategies based on interdependence.

Another suggestion for future work is studying the interaction between interdependence and violation severity on trust calibration. We speculate that more severe violations will result in higher trust decreases for teams engaged in joint actions. Future work could test these hypotheses by extending our research environment to include trust violations of different severity levels, such as machine failure during action execution and incorrect machine advice. These results could provide valuable insights for developing machines adapting to interdependence relationships to address trust calibration challenges.

5.3. Conclusion

Our study shows that interdependence relationships during human-machine teamwork influence human-machine trust calibration over time. During a simulated search and rescue task with a machine-induced trust violation, only in teams with joint actions does perceived trust in the machine not recover to its initial pre-violated value. However, our findings show that the correlation between perceived human-machine trust and machine trustworthiness is strongest in these teams with joint actions. This suggests that these stronger interdependence relationships during human-machine teamwork facilitate more accurate human-machine trust calibration. Overall, our study presents some first evidence that interdependence relationships during human-machine teamwork influence human-machine trust calibration over time. Therefore, it is crucial to consider these relationships carefully during human-machine trust calibration and to conduct follow-up research on adapting trust repair strategies to interdependence.

Acknowledgements This work is part of the research lab AI*MAN of Delft University of Technology.

References

[1] Verhagen RS, Neerincx MA, Tielman ML. Meaningful human control and variable autonomy in human-robot teams for firefighting. Frontiers in Robotics and AI. 2024 Feb;11. Available from: https://doi.org/10.3389/frobt.2024.1323980.

[2] Verhagen RS, Neerincx MA, Tielman ML. The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. Frontiers in Robotics and AI. 2022 Sep;9. Available from: https://doi.org/10.3389/frobt.2022.993997.

[3] Johnson M, Bradshaw JM, Feltovich PJ, Jonker CM, Van Riemsdijk MB, Sierhuis M. Coactive design: Designing support for interdependence in joint activity. Journal of Human Robot Interaction. 2014 Mar;3(1):43-69. Available from: https://doi.org/10.5898/JHRI.3.1.Johnson.

[4] Akata Z, Balliet D, De Rijke M, Dignum F, Dignum V, Eiben G, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer. 2020 Jul;53(8):18-28. Available from: https://doi.org/10.1109/MC.2020.2996587.

[5] Johnson M, Vera A. No AI is an island: the case for teaming intelligence. AI Magazine. 2019 Mar;40(1):16-28. Available from: https://doi.org/10.1609/aimag.v40i1.2842.

[6] Ososky S, Schuster D, Phillips E, Jentsch FG. Building appropriate trust in human-robot teams. In: 2013 AAAI Spring Symposium Series. 2013 Jan:60-65. Available from: https://cdn.aaai.org/ocs/5784/5784-24579-1-PB.pdf.

[7] Mehrotra S, Degachi C, Vereschak O, Jonker CM, Tielman ML. A systematic review on fostering appropriate trust in human-AI interaction. arXiv preprint. 2023 Nov. arXiv:2311.06305. Available from: https://doi.org/10.48550/arXiv.2311.06305.

[8] Lebiere C, Blaha LM, Fallon CK, Jefferson B. Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation. Frontiers in Robotics and AI. 2021 May;8. Available from: https://doi.org/10.3389/frobt.2021.652776.

[9] Kox ES, Kerstholt JH, Hueting TF, de Vries PW. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. Autonomous Agents and Multi-Agent Systems. 2021 Jun;35(2):30. Available from: https://doi.org/10.1007/s10458-021-09515-9.

[10] Kim PH, Ferrin DL, Cooper CD, Dirks KT. Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. Journal of Applied Psychology. 2004 Feb;89(1):104. Available from: https://psycnet.apa.org/doi/10.1037/0021-9010.89.1.104.

[11] Johnson M, Bradshaw JM. The role of interdependence in trust. Trust in Human-Robot Interaction. Academic Press. 2021 Mar:379-403. Available from: https://doi.org/10.1016/B978-0-12-819472-0.00016-2.

[12] Singh R, Sonenberg L, Miller T. Communication and shared mental models for teams performing interdependent tasks. In: Cranefield S, Mahmoud S, Padget J, Rocha A (eds) Coordination, Organizations, Institutions, and Norms in Agent Systems XII. COIN 2016. Lecture Notes in Computer Science (vol 10315). Springer, Cham. 2017 Aug;10315:81-97. Available from: https://doi.org/10.1007/978-3-319-46882-2_10.

[13] Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Academy of Management Review. 1995 Jul;20(3):709-734. Available from: https://doi.org/10.5465/amr.1995.9508080335.

[14] Jorge CC, Mehrotra S, Tielman ML, Jonker CM. Trust should correspond to trustworthiness: A formalization of appropriate mutual trust in human-agent teams. In: Falcone R, Zhang J, Wang D (Eds), Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021): Co-located with the 20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021). 2021 May;3022. Available from: https://pure.tudelft.nl/ws/portalfiles/portal/102967631/paper4.pdf.

[15] Schaubroeck J, Lam SS, Peng AC. Cognition-based and affect-based trust as mediators of leader behavior influences on team performance. Journal of Applied Psychology. 2011 Feb;96(4):863-871. Available from: https://psycnet.apa.org/doi/10.1037/a0022625.

[16] Jacovi A, Marasović A, Miller T, Goldberg Y. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Mar:624-635. Available from: https://doi.org/10.1145/3442188.3445923.

[17] Ferrario A, Loi M. How explainability contributes to trust in AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022 Jun:1457-1466. Available from: https://doi.org/10.1145/3531146.3533202.

[18] Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Human Factors. 2004 Feb;46(1):50-80. Available from: https://doi.org/10.1518/hfes.46.1.50_30392.

[19] Parasuraman R, Riley V. Humans and automation: Use, misuse, disuse, abuse. Human Factors. 1997 Jun;39(2):230-253. Available from: https://doi.org/10.1518/001872097778543886.

[20] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 2019 Feb;267:1-38. Available from: https://doi.org/10.1016/j.artint.2018.07.007.

[21] Verhagen RS, Neerincx MA, Tielman ML. A two-dimensional explanation framework to classify AI as incomprehensible, interpretable, or understandable. In: Calvaresi D, Najjar A, Winikoff M, Främling K (eds), Explainable and Transparent AI and Multi-Agent Systems. EXTRAAMAS 2021. Lecture Notes in Computer Science (vol 12688). Springer, Cham. 2021 Jul:119-138. Available from: https://doi.org/10.1007/978-3-030-82017-6_8.

[22] Zhang Y, Liao QV, Bellamy RK. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020 Jan:295-305. Available from: https://doi.org/10.1145/3351095.3372852.

[23] Tomsett R, Preece A, Braines D, Cerutti F, Chakraborty S, Srivastava M, et al. Rapid trust calibration through interpretable and uncertainty-aware AI. Patterns. 2020 Jul;1(4). Available from: https://doi.org/10.1016/j.patter.2020.100049.

[24] Chen JY, Lakhmani SG, Stowers K, Selkowitz AR, Wright JL, Barnes M. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theoretical Issues in Ergonomics Science. 2018 Feb;19(3):259-282. Available from: https://doi.org/10.1080/1463922X.2017.1315750.

[25] Mercado JE, Rupp MA, Chen JY, Barnes MJ, Barber D, Procci K. Intelligent agent transparency in human–agent teaming for multi-UxV management. Human Factors. 2016 Feb;58(3):401-415. Available from: https://doi.org/10.1177/0018720815621206.

[26] Selkowitz AR, Lakhmani SG, Chen JY. Using agent transparency to support situation awareness of the Autonomous Squad Member. Cognitive Systems Research. 2017 Dec;46:13-25. Available from: https://doi.org/10.1016/j.cogsys.2017.02.003.

[27] De Visser EJ, Peeters MM, Jung MF, Kohn S, Shaw TH, Pak R, Neerincx MA. Towards a theory of longitudinal trust calibration in human–robot teams. International Journal of Social Robotics. 2020 May;12(2):459-478. Available from: https://doi.org/10.1007/s12369-019-00596-x.

[28] Guo Y, Yang XJ. Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach. International Journal of Social Robotics. 2021 Dec;13(8):1899-1909. Available from: https://doi.org/10.1007/s12369-020-00703-3.

[29] Lewicki RJ, Brinsfield C. Trust repair. Annual Review of Organizational Psychology and Organizational Behavior. 2017 Jan;4:287-313. Available from: https://doi.org/10.1146/annurev-orgpsych-032516-113147.

[30] Esterwood C. Rethinking trust repair in human-robot interaction. In: Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing. 2023 Oct:432-436. Available from: https://doi.org/10.1145/3584931.3608919.

[31] Esterwood C, Robert LP Jr. Three strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. Computers in Human Behavior. 2023 May;142. Available from: https://doi.org/10.1016/j.chb.2023.107658.

[32] Kramer RM, Lewicki RJ. Repairing and enhancing trust: Approaches to reducing organizational trust deficits. The Academy of Management Annals. 2010 Jun;4(1):245-277. Available from: https://doi.org/10.1080/19416520.2010.487403.

[33] Schweitzer ME, Hershey JC, Bradlow ET. Promises and lies: Restoring violated trust. Organizational Behavior and Human Decision Processes. 2006 Sep;101(1):1-19. Available from: https://doi.org/10.1016/j.obhdp.2006.05.005.

[34] Esterwood C, Robert LP. A literature review of trust repair in HRI. In: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE; 2022 Sep:1641-1646. Available from: https://doi.org/10.1109/RO-MAN53752.2022.9900667.

[35] Robinette P, Howard AM, Wagner AR. Timing is key for robot trust repair. In Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7.

Springer International Publishing; 2015 Oct:574-583. Available from: https://doi.org/10.1007/978-3-319-25554-5_57.

[36] Perkins R, Khavas ZR, McCallum K, Kotturu MR, Robinette P. The reason for an apology matters for robot trust repair. In: International Conference on Social Robotics. Cham: Springer Nature Switzerland; 2023 Feb:640-651. Available from: https://doi.org/10.1007/978-3-031-24670-8_56.

[37] Xu J, Howard A. Evaluating the impact of emotional apology on human-robot trust. In: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE; 2022 Sep:1655-1661. Available from: https://doi.org/10.1109/RO-MAN53752.2022.9900518.

[38] Sebo SS, Krishnamurthi P, Scassellati B. "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). IEEE; 2019 Mar:57-65. Available from: https://doi.org/10.1109/HRI.2019.8673169.

[39] Hoffman RR, Mueller ST, Klein G, Litman J. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science. 2023 Feb;5. Available from: https://doi.org/10.3389/fcomp.2023.1096257.

[40] Noguchi K, Gel Y, Brunner E, Konietschke F. nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. Journal of Statistical Software. 2012 Sep;50. Available from: https://doi.org/10.18637/jss.v050.i12.

[41] De Visser EJ, Monfort SS, McKendrick R, Smith MA, McKnight PE, Krueger F, Parasuraman R. Almost human: Anthropomorphism increases trust resilience in cognitive agents. Journal of Experimental Psychology: Applied. 2016 Aug;22(3):331-349. Available from: https://psycnet.apa.org/doi/10.1037/xap0000092.

[42] Collins MG, Juvina I. Trust miscalibration is sometimes necessary: An empirical study and a computational model. Frontiers in Psycholgy. 2021 Aug;12. Available from: https://doi.org/10.3389/fpsyg.2021.690089.

[43] Verhagen RS, Neerincx MA, Parlar C, Vogel M, Tielman ML. Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance. In: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems. 2023 May;2316-2318. Available from: https://pure.tudelft.nl/ws/portalfiles/portal/155562598/3545946.3598919.pdf.