Trust Repair in Human-Agent Teams: the Effectiveness of Explanations and Expressing Regret

JAAMAS Track

E.S. Kox
TNO
Soesterberg, The Netherlands
esther.kox@tno.nl

J.H. Kerstholt
TNO
Soesterberg, The Netherlands
jose.kerstholt@tno.nl

T.F. Hueting
TNO
Soesterberg, The Netherlands
tom.hueting@tno.nl

P.W. de Vries
University of Twente
Enschede, The Netherlands
p.w.devries@utwente.nl

ABSTRACT

Trust is a fundamental aspect of teamwork in Human-Agent Teams (HATs). Trust violations are an inevitable aspect of the cycle of trust, so effective trust repair strategies are needed to ensure durable and successful team performance. This study explores the effectiveness of four trust repair strategies. In a first-person shooter resembling HAT task, a trust violation was provoked when the robotic agent failed to detect an approaching enemy. After this, the agent offered an apology composed of an explanation and/or an expression of regret (either one alone, both or neither). Our results indicated that expressing regret was crucial for effective trust repair, and that trust repair was most effective when the apology contained both components.

KEYWORDS

Trust; Human-Agent Teaming; Trust Repair

ACM Reference Format:
E.S. Kox, J.H. Kerstholt, T.F. Hueting, and P.W. de Vries. 2022. Trust Repair in Human-Agent Teams: the Effectiveness of Explanations and Expressing Regret: JAAMAS Track. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

Humans and autonomous agents are increasingly accomplishing goals together, like driving cars and performing surgery as Human-Agent Teams (HATs). We define a HAT as a team consisting of at least one human and one intelligent agent, robot, and/or other AI or autonomous system. The artificial component of the team will be referred to as an autonomous agent (AA), defined as an artificial entity that observes and acts upon an environment autonomously and that is able to communicate and collaborate with other agents, including humans, to solve problems and achieve (common) goals.

Trust is a fundamental aspect of such teamwork. The highly interdependent and dynamic nature of teamwork demands trust among team members to be correctly calibrated in order to perform successfully. Trust is correctly calibrated when the level of human trust is warranted by the agent's capabilities [7]. If the former exceeds the latter, this may cause humans to overly rely on the agent ("overtrust"); the latter exceeding the former may result in disuse ("undertrust"). Both can lead to inappropriate reliance on AAs, which can compromise safety and profitability [7]. Trust is defined as the human's willingness to make oneself vulnerable and to act on the AA's recommendations and decisions in the pursuit of some benefit, with the expectation that the AA will help achieve their common goal in an uncertain context [6]. Especially under complex and risky conditions, the establishment of calibrated trust among teammates is essential for efficient collaboration and communication [3, 9].

Given the complexity and unpredictability of many real-world situations, it is inevitable that the AA will err at some point. This may lead to a decrease in both trust in the AA and willingness to accept further information from the AA, leading to a limited benefit from the advantages that AAs have to offer [4]. This underscores the importance of effective trust repair strategies.

The current study focuses on apology as a trust repair strategy [8]. An apology can consist of multiple components, including an expression of regret and an explanation. Generally, research shows that providing an apology can benefit the feelings of the human towards an artificial entity [1, 2, 11]. The effectiveness of a trust repair strategy seems to depend on the composition of the apology [8] as well as on situational factors, like timing [10] and agent type [5]. New approaches are needed to understand the potential impact of apologetic messages from non-human agents on human-agent trust.

The aim of this study is to investigate the effect of different apology compositions on the repair of trust after a trust violation in a HAT context. We expect to find an effect for both components alone (expression of regret and explanation), but the combination of components is expected to be the most effective trust repair strategy.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

2 METHOD

2.1 Design

A 3 (Time: prior to violation [T1], after violation [T2], after repair [T3]) x 2 (Regret: provided or not) x 2 (Explanation: provided or not) mixed-design was used. Time was a within-participant factor and Regret and Explanation were varied between participants. The main dependent variable was Trust. Participants (n=66) were randomly assigned to one of the four trust-repair conditions (explanation only: n=18; regret only: n=16; neither: n=14; both: n=18).

2.2 Task and procedure

Participants carried out a mission in a first-person shooter resembling environment, with a robotic character as their teammate (the AA). For this, the Wizard of Oz method was used: the AA was controlled by an experiment leader in an adjacent room, while the participant was kept under the impression that it was operating autonomously.

Participants were instructed to head back to basecamp as fast and careful as possible. Meanwhile, the AA reported whether it detected enemies or not and provided the corresponding advice to take shelter or continue moving (via audio messages). After an advice, the game paused and participants rated their willingness to accept the AA's advice. After that participants heard whether the previous advice had been correct or not. Feedback was either provided auditorily (e.g. "my advice was correct"), or by an external event (i.e. the appearance of an enemy). This happened after the AA's second advice ("I do not detect any danger"). This incorrect advice is meant to provoke a trust violation. After receiving feedback, participants were asked to rate their trust in the AA. Trust was measured thrice (prior to violation [T1], after violation [T2], after repair [T3]). The trust repair manipulation followed the second trust measure. A schematic timeline is presented in Figure 1.

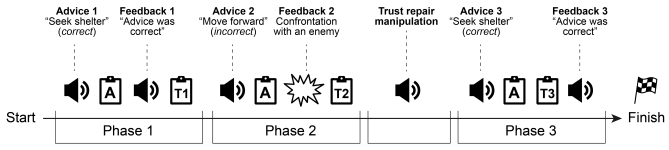

Figure 1: Schematic timeline of the experiment

2.3 Manipulation

The trust repair message varied between participants as it depended on the manipulated factors Explanation and Regret. Every trust repair strategy, including baseline condition, started with an acknowledgement "The advice I gave you was wrong". The explanation was: "The enemy was carrying a weapon of an ally, because of that, my classification led to an incorrect conclusion". In the conditions with an expression of regret, the agent would end its message with "I am really sorry".

3 RESULTS

A significant main effect for Time [T1-T3] on Trust was obtained with $F(2,124) = 53.66$, $p < .001$. Results of the LSD post-hoc test shows a significant difference between T1 (M = 5.06) and T2 (M = 4.01) ($p < .001$), which reflects a successful violation of trust. Moreover, a significant difference between T2 and T3 (M = 4.44) ($p < .001$) was found, which reflects an overall trust recovery effect.

A significant three-way interaction effect between Time [T1-T3], Explanation and Regret on trust was found with $F(2, 124) = 3.31$, $p = .040$. LSD post-hoc analysis shows a significant difference between groups in how they react to the incorrect advice prior to T2. On average, the participant group in the condition with both regret and explanation shows significantly lower levels of trust at T2 compared to participants groups in the conditions with solely explanation ($p = .007$) and the condition with solely regret ($p = .010$) at T2.

To measure the effects of the trust repair strategies, simple effects were calculated to compare trust scores before and after provision [T2-T3] for each experimental condition. Increases in trust between T2 and T3 were only significant when an expression of regret was provided. This effect is marginally significant when no explanation is given ($p = .056$), and stronger when it is accompanied by an explanation ($p < .001$). The effect is non-significant when the agent provides only an explanation ($p = .142$) or neither components ($p = .199$).

4 DISCUSSION

The results of this study show that the trust repair strategies including an expression of regret (i.e., "I am sorry") were most effective in repairing trust after a trust violation in a human-agent teaming setting. Trust was only significantly recovered when the apology included an expression of regret. This effect was strongest in combination with an explanation. Although expressing regret is typically perceived as a human-like quality, these results suggest that saying sorry can also makes a difference in rebuilding trust when it comes from a non-human agent. As AAs are increasingly deployed as in more social roles, it seems useful to incorporate social cues like apologies into their design. Even though the technology evolves at a high rate, we must prepare for the inevitability of errors. This study contributes to the exploration of strategies for the maintenance and repair of trust in human-agent teaming. To retain trust in a human-agent team, the ability of actively repairing trust after an error or unintended action should be a fundamental part of the design of AAs.

ACKNOWLEDGMENTS

Please read the journal version of this article for more details. This material is based upon work supported by the Dutch Ministry of Defense's exploratory research program

REFERENCES

[1] M. Akgun, K. Cagiltay, and D. Zeyrek. 2010. The effect of apologetic error messages and mood states on computer users' self-appraisal of performance. *J. Pragmat* 42, 9 (2010), 2430–2448.

[2] S. Brave, C. Nass, and K. Hutchinson. 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human Computer Studies* 62, 2 (2005), 161–178.

[3] E.J. de Visser et al. 2019. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics* (2019), 1–20.
[4] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. De Visser, and R. Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53, 5 (2011), 517–527.
[5] T. Kim and H. Song. 2021. How should intelligent agents apologize to restore trust?: The interaction effect between anthropomorphism and apology attribution on trust repair. *Telemat. Informatics* 53 (2021).
[6] Esther S. Kox, José H. Kerstholt, Tom F. Hueting, and Peter W. de Vries. 2021. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems* 35, 2 (2021), 1–20.
[7] J. D. Lee and K. A. See. 2021. Trust in Automation: Designing for Appropriate Reliance. 46, 1 (2021), 50–80.
[8] R. J. Lewicki, B. Polin, and R. B. Lount. 2016. An Exploration of the Structure of Effective Apologies. *Negot. Confl. Manag. Res.* 9, 2 (2016), 177–196.
[9] R. Parasuraman, E. De Visser, E. Wiese, and P. Madhavan. 2014. Human trust in other humans, automation, robots, and cognitive agents: Neural correlates and design implications. *Proc. Hum. Factors Ergon. Soc.* 14, 1 (2014), 340–344.
[10] P. Robinette, A.M. Howard, and A.R. Wagner. 2015. Timing is key for robot trust repair. *International conference on social robotics* 9388 (2015), 574–583.
[11] J. Y. Tzeng. 2004. Toward a more civilized design: Studying the effects of computers that apologize. *International Journal of Human Computer Studies* (2004).