CSE3000 Research project: Agent Failure, Trust Repair, and Fluency in Human-AI Team

~ Impact of Opportunistic Interdependence Relationship in a Human-Agent Team ~

Author: Kanta Tanahashi Supervisor: Ruben Verhagen Responsible Professor: Myrthe Tielman Contact: k.tanahashi@student.tudelft.nl

1. Introduction

Human Autonomous Teams (HATs) combine capabilities to perform tasks more efficiently.

Trust recovery after **trust violations** is important to maintain a high trust level, which is crucial to team performance.

Past studies found:

- Communicating **uncertainty** in advice **mitigates** trust loss following trust violation [1].
- Expressing **regret/**providing **explanations** in apology is an effective trust repair strategy [2].

Collaborative fluency: measurement of coordination and meshing of actions in a team[3].

Interdependence relationship: set of complementary relationships that parties rely on to manage joint activities [4].

2. Research Question

How do **opportunistic (soft) interdependence relationships** affect

1) **trust violation** and **trust repair**

2) **collaboration fluency**

compared against independence (baseline) condition?

Why soft interdependence?

- It is necessary to achieve true teamwork [5].
- Successful teams tend to manage soft interdependence well [5].

3. Methodology

MATRX was used to conduct a user study.

Objective: to collaborate with an AI agent (RescueBot) and to rescue victims in different areas.

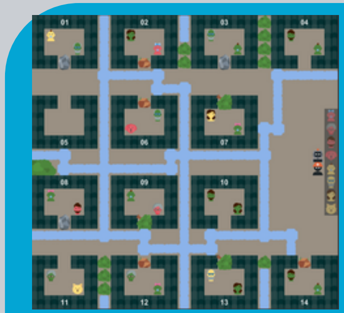
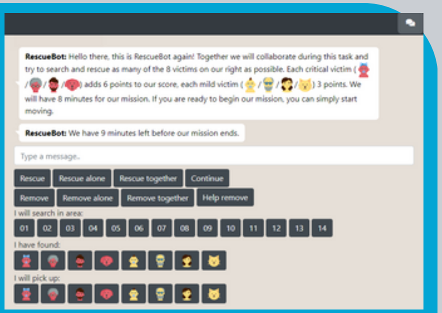


Figure 1: God view of the environment


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.

RescueBot: We have 9 minutes left before our mission ends.

Figure 2: Messaging functionality

Procedure:

During the game, three extreme rains arrived. Getting hit by rain led to reduction in playing time/score. Before each rain, weather forecasting message was sent by RescueBot:

- ***1st advice:*** at 2 minutes mark. **Correctly** predicts the storm.
- **2nd advice**: at 4 minutes mark. **Incorrectly** predicts light rain, leading to trust violation. Trust repair follows.
- ***3rd advice:*** at 6 minutes mark. **Correctly** predicts the storm.

Trust was measured after the three advice/feedback with a questionnaire based on the trust scale for XAI context [6]. Collaboration fluency was measured with existing questionnaire [3] and objective metrics.

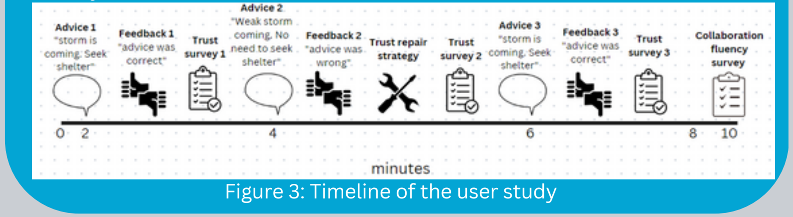


Figure 3: Timeline of the user study

4. Results and discussion

Trust:

- Significantly higher trust for opportunistic condition before trust violation (T1)

↓

Influenced by the sense of team structure

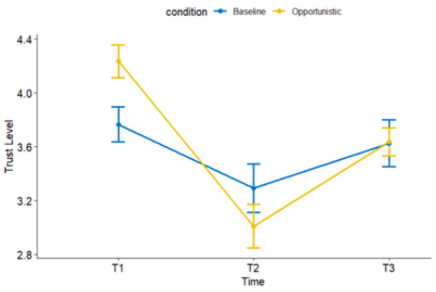


Figure 4: Effect of opportunistic interdependence (yellow) and baseline (blue) on trust at different times

- Opportunistic experienced significant trust violation/recovery

↓

Interdependence supports continuous calibration of trust

Collaboration Fluency:

- No significant difference based on questionnaire
- Objective metrics showed a difference in ratio of robot idle time but nothing significant

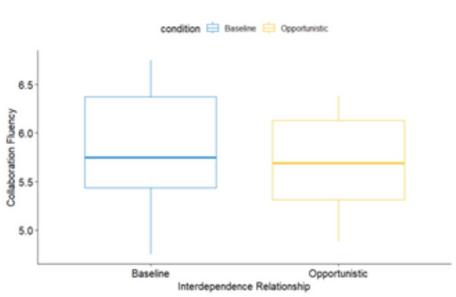


Figure 5: Analyzing the effect of opportunistic interdependence (yellow) and baseline (blue) on fluency

Limitations in the experiment:

- Arrow key press speed
- Difference in trust and fluency between subjects who got punished by rain/who did not etc

5. Conclusions and future work

- Trust was higher for opportunistic due to team structure
- Trust violation/recovery more affected by opportunistic due to role of interdependence to support continuous exploration of trust
- No significant result in terms of fluency

Possible next step:

Investigate a type of trust repair strategy that is effective for the teams with opportunistic interdependence in particular.

References

[1] 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)

[2] 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.

[3]Guy Hoffman. “Evaluating Fluency in HumanâRobot Collaboration”. In: IEEE Transactions on Human-Machine Systems 49.3 (2019), pp. 209–218. doi: 10.1109/THMS.2019.2904558.

[4] Matthew Johnson, Jeffrey Bradshaw, Paul J. Feltovich, Catholijn Jonker, M. Riemsdijk, and Maarten Sierhuis. Coactive design: Designing support for interdependence in joint activity. Human Robot-Interaction, 3(1), 03 2014

[5] 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.

[6] 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.