A Poster Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
May 16, 2023
Final project course: CSE3000 Research Project

EEMCS, Delft University of Technology, The Netherlands

AGENT FAILURE AND TRUST REPAIR IN HUMAN-AGENT TEAMS

AUTHOR

Tauras Narbutas
T.Narbutas@student.tudelft.nl

COLLABORATORS

Cherin Kim, Alexandra Marcu, Kanta Tanahashi

SUPERVISORS

Myrthe Tielman, Ruben Verhagen

EXAMINER

Ujuwal Gadiraju

01 BACKGROUND

Collaborative AI

- Human-agent teams rely on interdependence relationships, meaning that both parties have to work together on certain sub-tasks in order to achieve a common goal [1, 2, 3]

Trust

- Prior research shows that expressing regret and providing an explanation are effective trust-repair strategies [4]

Collaboration fluency

- Collaboration fluency investigates how smoothly and efficiently human-agent teams interact and work together toward achieving common tasks

02 OBJECTIVES

How does the **required** interdependence relationship arising from a lack of human and robot capacities affect

1. **the trust violation**
2. **the trust repair**
3. **the collaboration fluency**

in human-agent teams compared to the **baseline condition** where individuals cooperate independently ?

03 METHODOLOGY

User study

- 30 participants - 15 for baseline and 15 for required conditions respectively
- The questionnaires were created using the Qualtrics tool
- The game dynamics were implemented using the human-agent teaming rapid experimentation software package MATRX
- The task Search and rescue mission in a town affected by extreme weather (heavy rain) and floods
- Collaborative efforts were needed between the human participant and the AI agent
- The goal was to save 4 critically (6 points) and 4 mildly (3 points) injured victims while removing areas blocking objects

Measures

1. Subjective
   a. Trust questionnaire
   b. Collaboration fluency questionnaire
2. Objective
   a. Performance (completeness, score, task duration)
   b. Agent idle time
   c. Number of human-sent messages
   d. Human location during storm

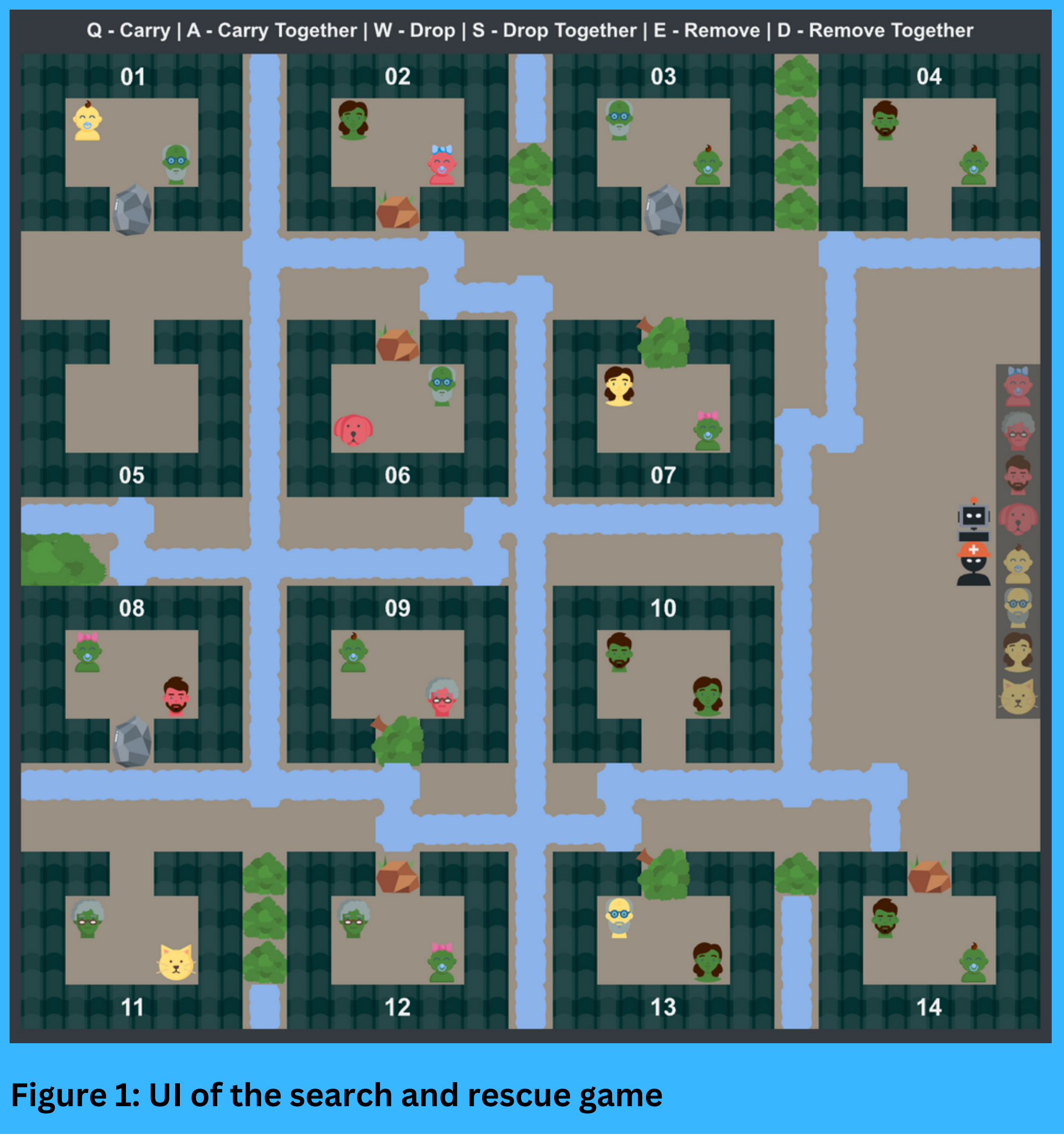
Q - Carry | A - Carry Together | W - Drop | S - Drop Together | E - Remove | D - Remove Together
01
02
03
04
05
06
07
08
09
10
11
12
13
14

Figure 1: UI of the search and rescue game

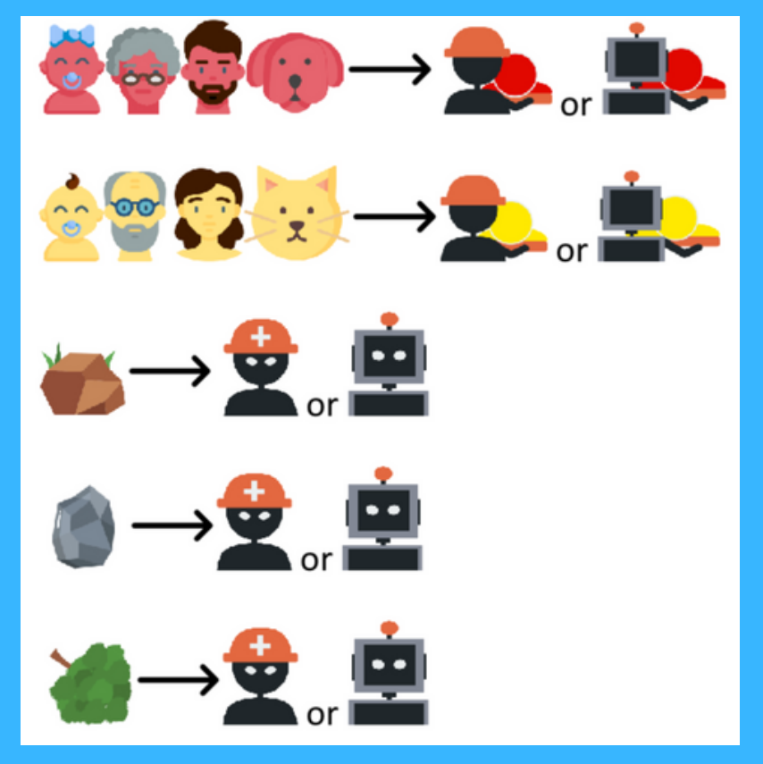

Figure 2: Baseline condition

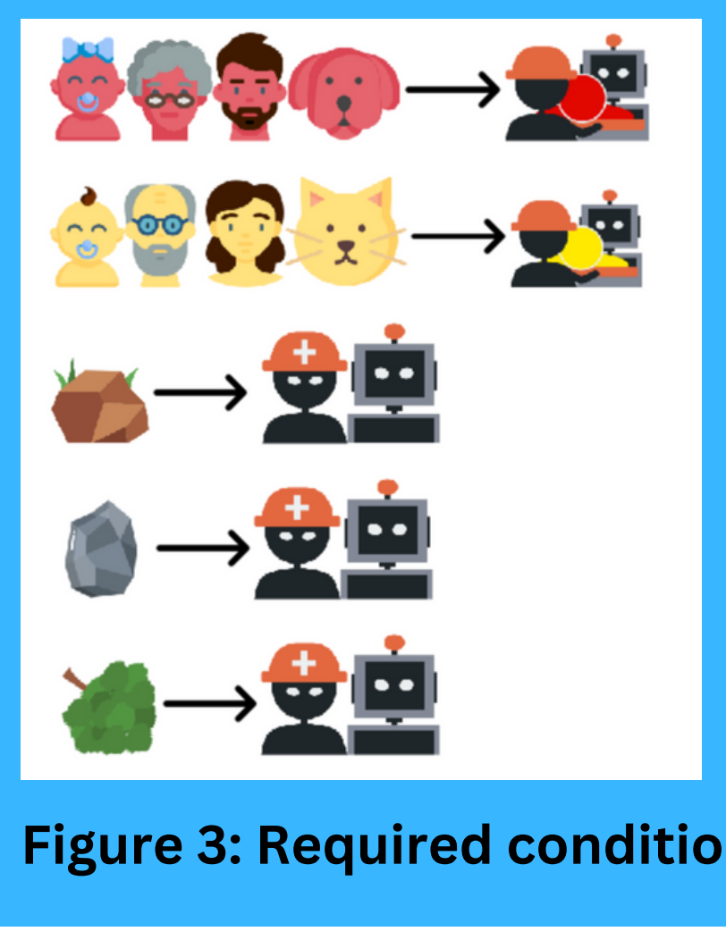

Figure 3: Required condition

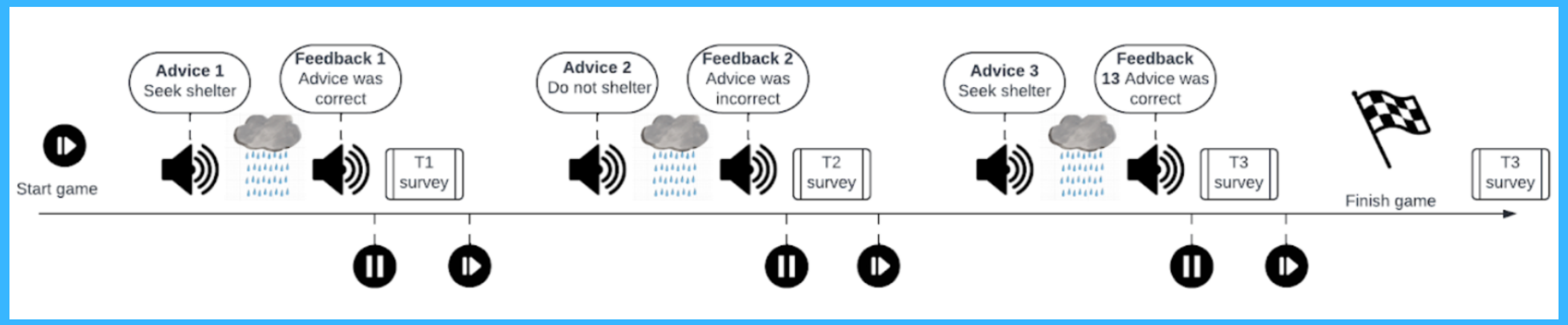

Figure 4: Schematic timeline depicting the user study

04 RESULTS

1. Subjective measures
   a. **Required condition** resulted in both **higher trust violation** and **higher trust repair** (Figure 5)
   b. **No significant effect** was found on **collaboration fluency**
2. Objective measures
   a. **Baseline condition** resulted in **higher completeness** and **lower time duration** (Figures 6 and 7 respectively), **no effec**t was found on the **score**
   b. **Required interdependence** resulted in **higher agent idle time** (Figure 8)
   c. **No significant effect** was found on the **number of human-sent message**s
   d. **1.5 times more** participants hid from heavy rain when advised by the agent for the **required condition**

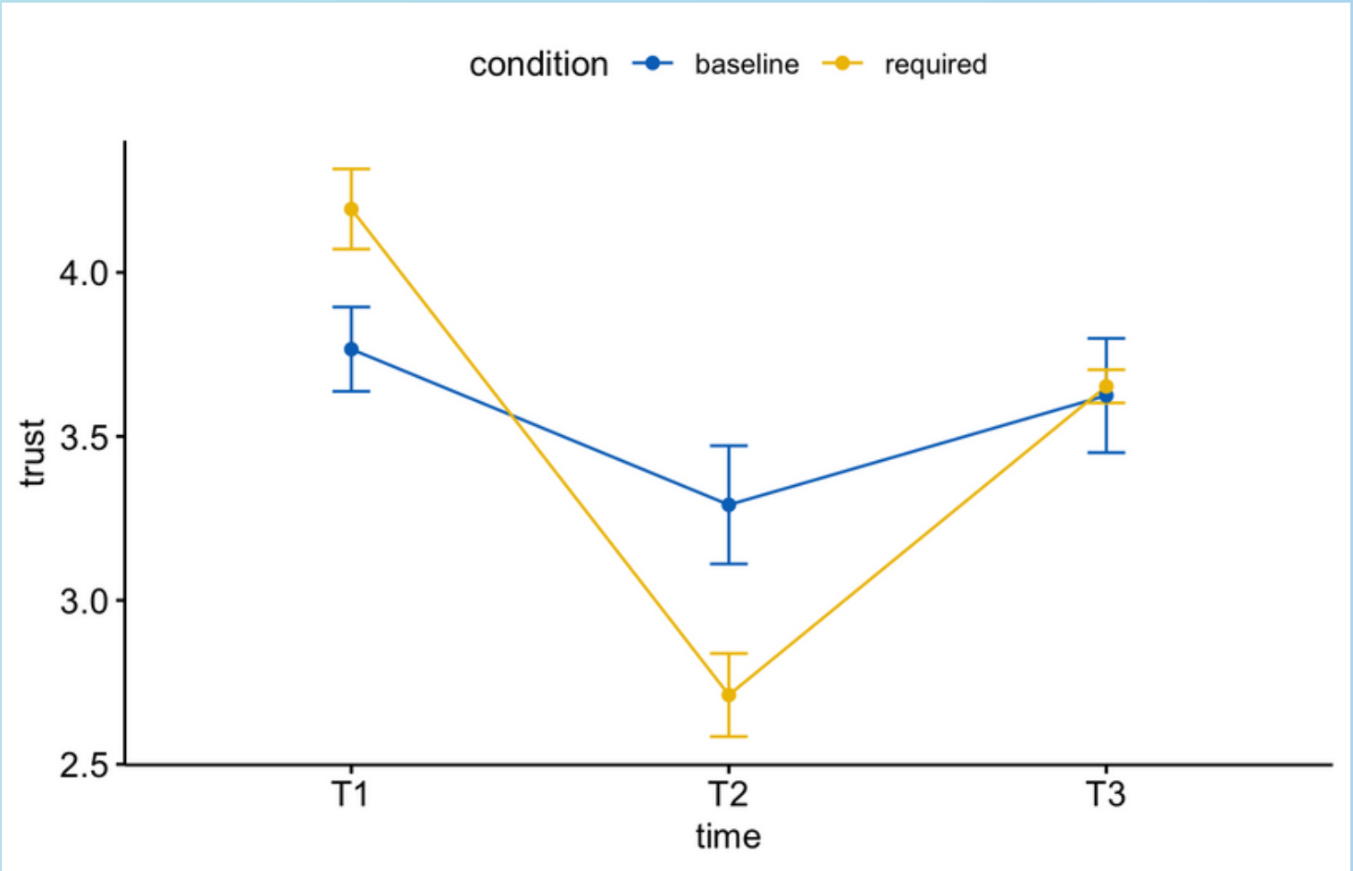

Figure 5: Estimated marginal means illustrating the relationship of trust and time per condition

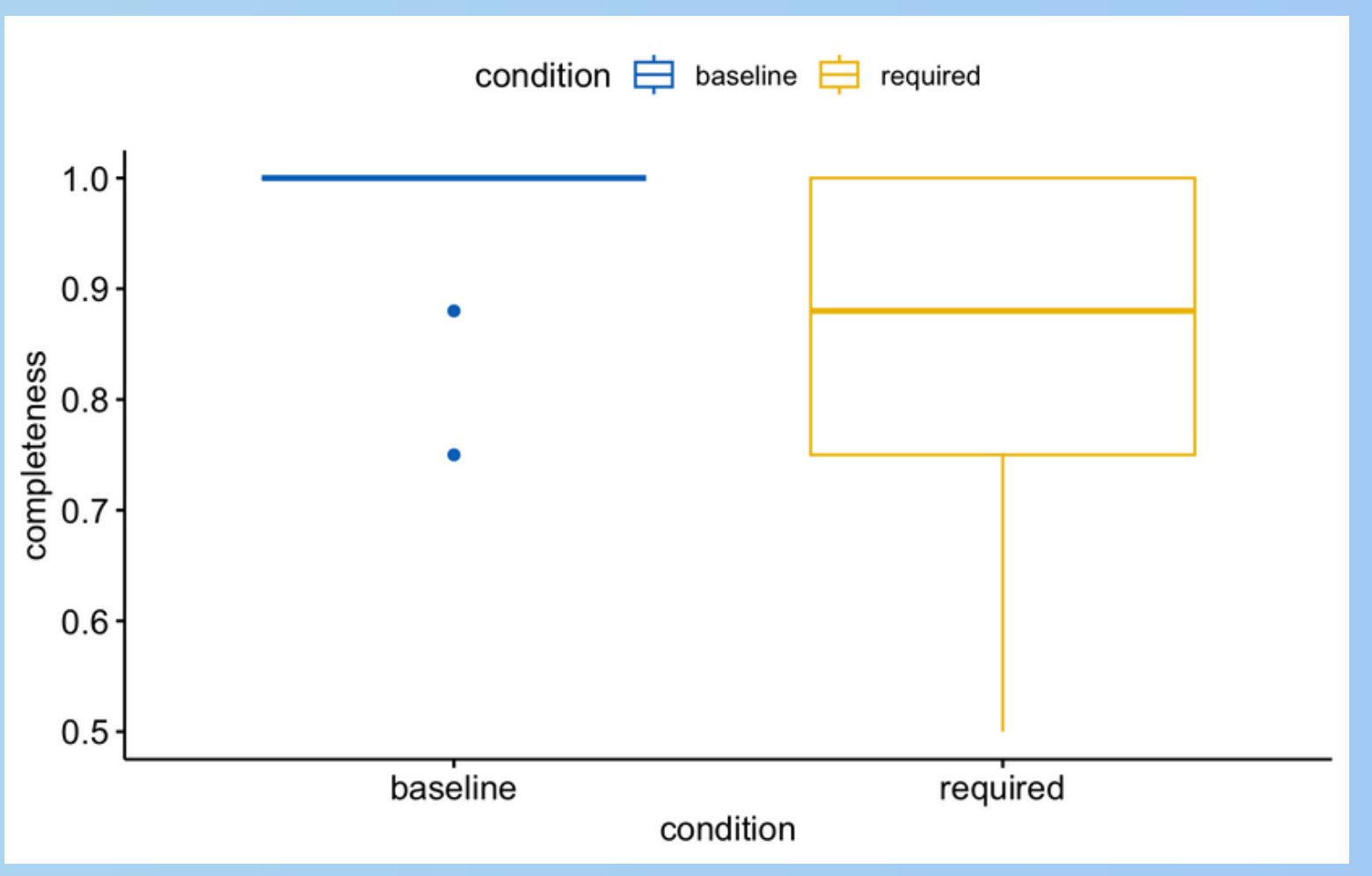

Figure 6: Box plot depicting the completeness of the game per condition

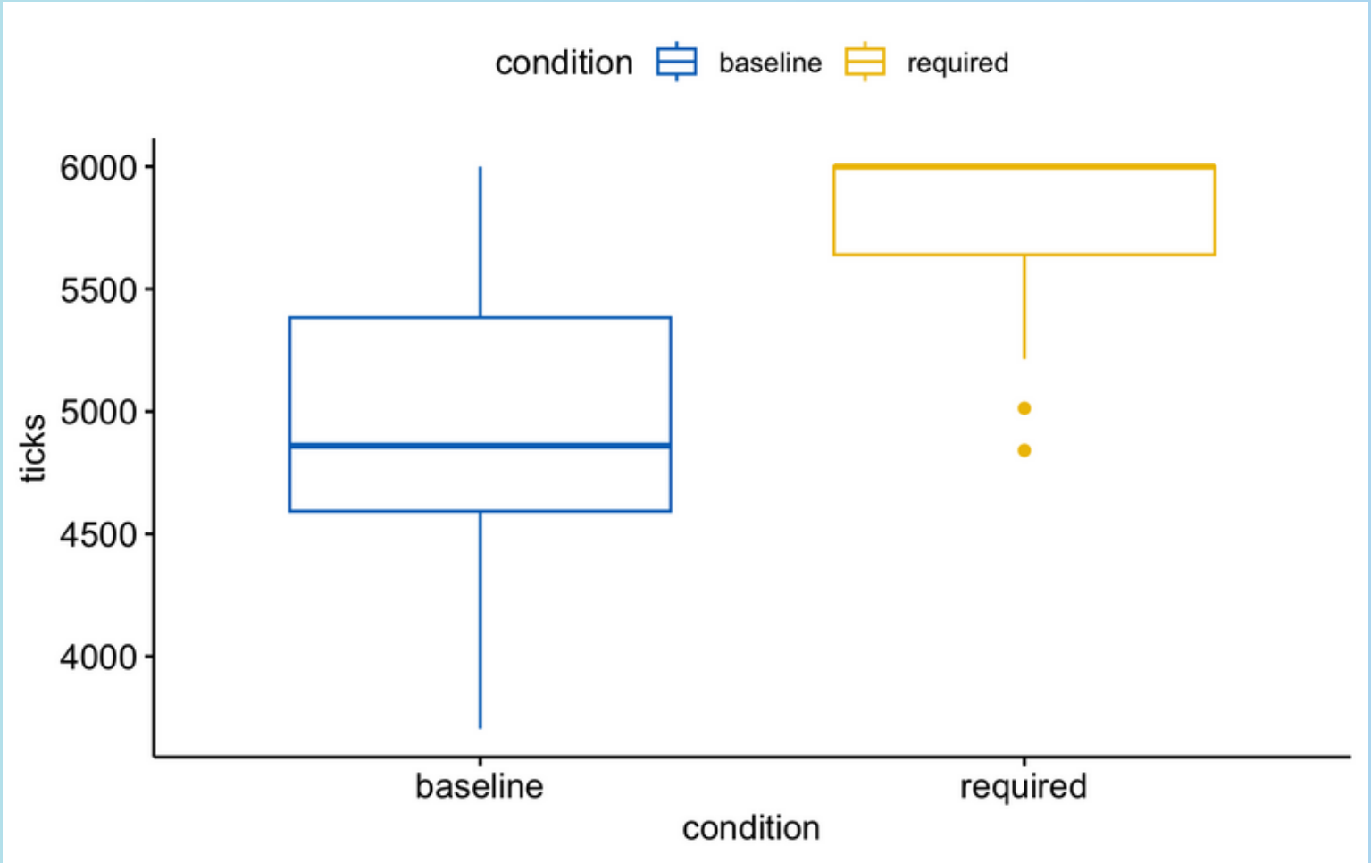

Figure 7: Box plot depicting the time needed to finish the task per condition

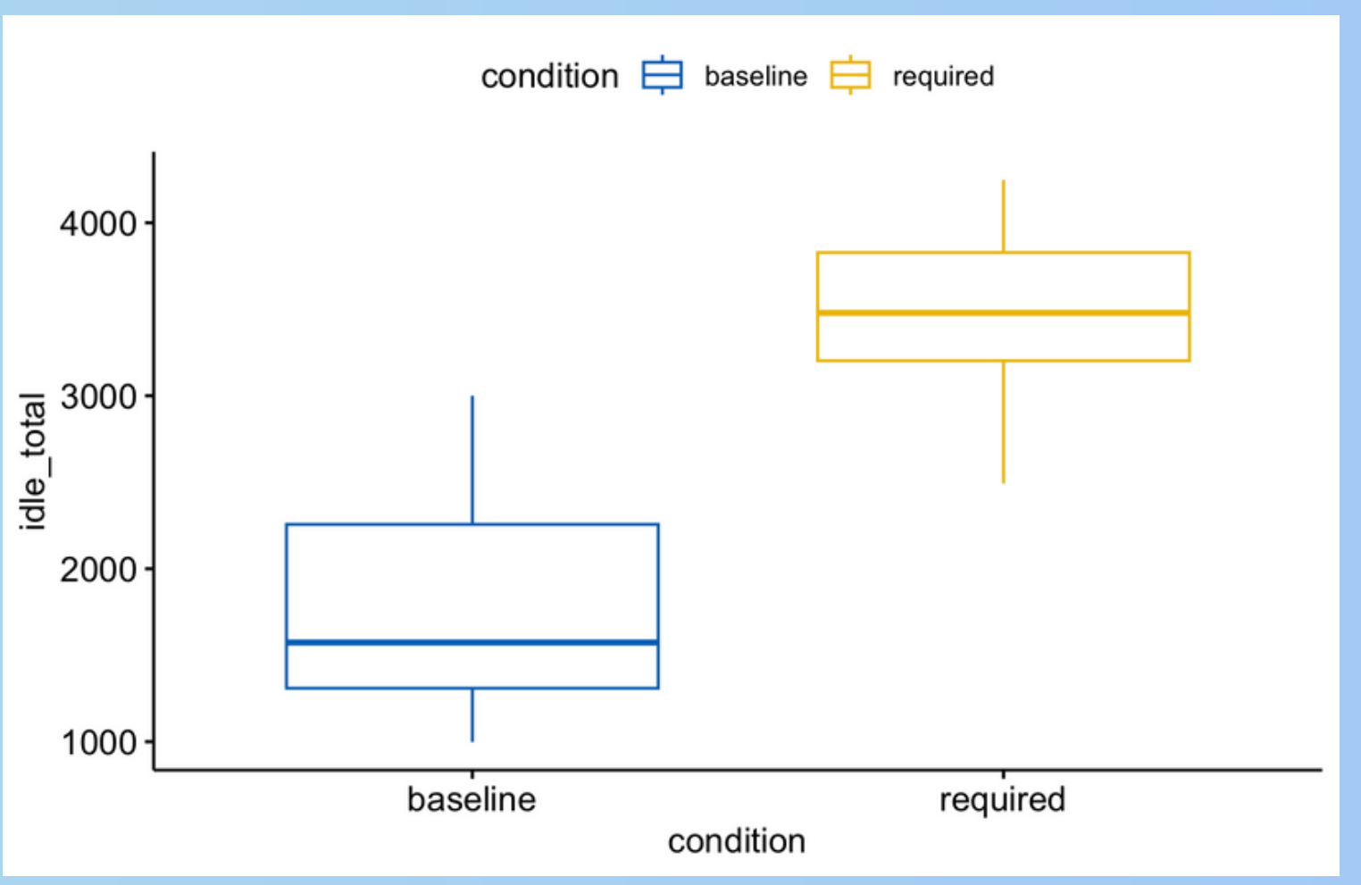

Figure 8: Box plot depicting AI agent idle time per condition

06 CONCLUSIONS

Trust violation and trust repair

- Interdependence condition had a significant effect on both trust violation and trust repair as assessed by the questionnaire results of the user study

Collaboration fluency

- Interdependence condition had no significant effect on collaboration fluency as assessed by the questionnaire results of the user study
- Objective measures including completeness, time taken to complete the task, and agent idle time imply more effective collaboration for the baseline condition
- For the required condition, human likeliness to follow agent advice implies better team communication

REFERENCES

[1] Johnson, M., & Vera, A. (2019). No AI is an island: the case for teaming intelligence. AI magazine, 40(1), 16-28.
[2] Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. Journal of Human-Robot Interaction, 3(1), 43-69.
[3] Verhagen, R. S., Neerincx, M. A., & Tielman, M. L. (2022). The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. Frontiers in Robotics and AI, 9, 243.
[4] Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. Autonomous agents and multi-agent systems, 35(2), 1-20.