COMMUNICATING TRUST-BASED BELIEFS AND DECISIONS IN HUMAN-AI TEAMS USING VISUAL SUMMARIES OF EXPLANATIONS

AUTHOR: SAHAR MAROSSI
SUPERVISOR: CAROLINA FERREIRA GOMES CENTEIO JORGE
RESPONSIBLE PROFESSOR: MYRTHE TIELMAN

1. INTRODUCTION & BACKGROUND

Human-AI Teams: HATS - [1]

Teams of both a **human** and an **artificial agent** working towards a team **goal**, typically composed of a **set of tasks** that can be performed either individually or jointly.

Trust - [2], [3]

Dyadic behavior between a **trustor** and a **trustee**. The "**willingness**" of one party to be open to the **risks** posed by another party's **actions**.

- **Artificial trust:** Artificial agents trusting humans.
- **Natural trust:** Humans trusting artificial agents.

Mental models - [4], [5]

Structured **mental representations** to describe, explain, and predict the surrounding **environment**.

- To ensure trust, **communication** is key. This can be done by **sharing** the agent's **mental model**.
- This facilitates a **feedback loop** (Figure 1).

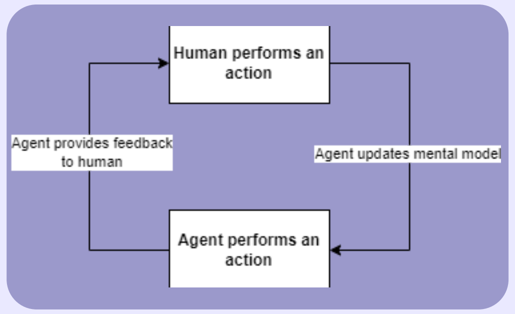

Human performs an action
Agent updates mental model
Agent performs an action
Agent provides feedback to human

Figure 1: Feedback loop of updating trust and communication of mental models

2. RESEARCH QUESTIONS

How does a **visual summary of explanations** of the **mental model** of the agent's **trust** (artificial trust) in the human teammate affect:

- **RQ1:** The human teammate's trust in the agent (natural trust)?
- **RQ2:** The human teammate's overall satisfaction in the agent?

3. TRUST MODEL & EXPERIMENT

Environment

The **human (user)** and the **agent (RescueBot)** are given the mission of ***searching/rescuing*** victims in an urban search and rescue environment (Figure 2). Tasks include **searching rooms**, **removing obstacles**, and **rescuing victims.**

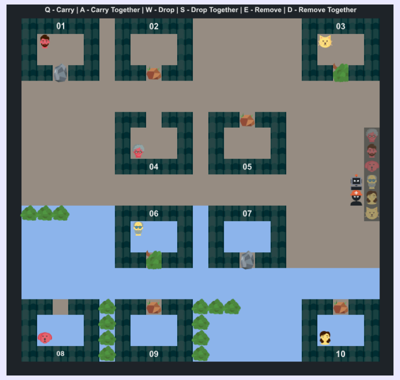

Q - Carry | A - Carry Together | W - Drop | S - Drop Together | E - Remove | D - Remove Together
01
02
03
04
05
06
07
08
09
10

Figure 2: Image of the environment (map) in God view

Trust Model

RescueBot will have a **mental model** of its **trust beliefs** regarding the human teammate's **competence** and **willingness**. This model will influence the agent's behavior and decisions.

Visual Summary

- Time series plot of **trust beliefs** vs time (Figure 3).
- **Interactive data points.** Hovering over them displays an explanation for the change in trust beliefs.
- **Verdict** explaining the agent's **behavior** and future **decisions**.

Experiment

The **task** was to **rescue 6 victims** (3 mild, 3 critical) within **10 minutes**. **Two conditions** were compared:

- **Baseline** (no visual summaries).
- **Summary** (visual summaries were shown **3 times** throughout the task).

Measures

Subjective Measures are measured with questionnaires**:**

- Natural Trust.
- Satisfaction.

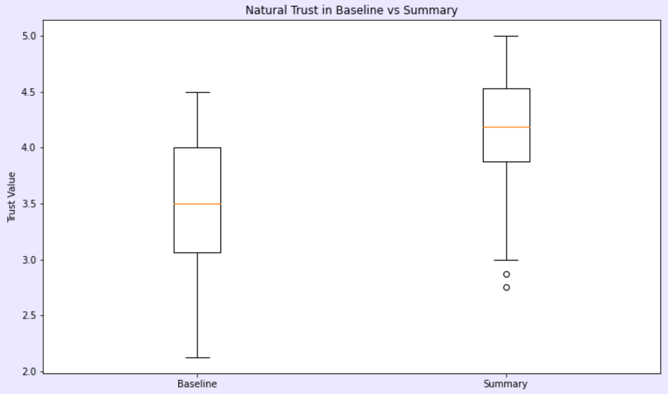
Objective Measures are logged automatically:

- Artificial Trust (average competence & willingness).

4. RESULTS

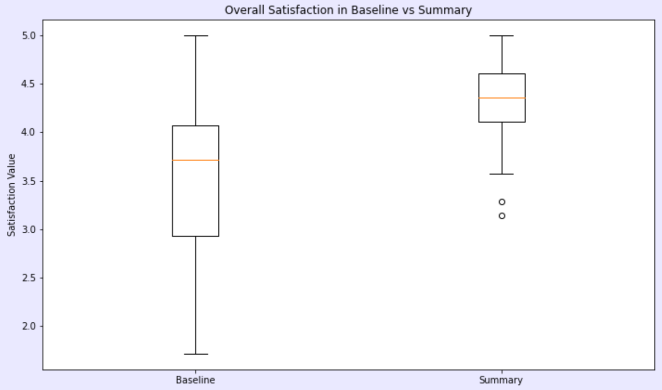
Natural Trust

- Shapiro-Wilk tests succeed on both datasets.
- Levene's test succeeds → t-test performed.
- Significant difference found (p = 0.0028).



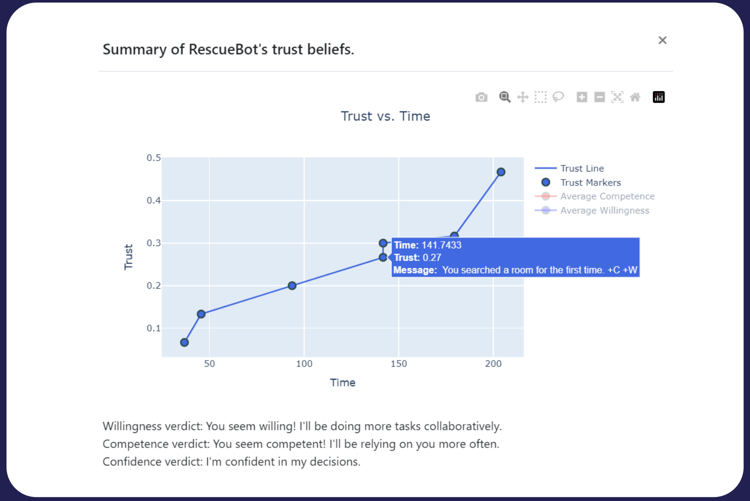
Overall Satisfaction

- Shapiro-Wilk tests succeed on both datasets.
- Levene's test fails → Welch t-test performed.
- Significant difference found (p = 0.0034).



Performance

- Statistical significance found for **Artificial Trust** (summary > baseline).



Summary of RescueBot's trust beliefs.

Time: 141.7433
Trust: 0.27
Message: You searched a room for the first time. +C +W

Willingness verdict: You seem willing! I'll be doing more tasks collaboratively.
Competence verdict: You seem competent! I'll be relying on you more often.
Confidence verdict: I'm confident in my decisions.

Figure 3: Visual summary of the agent's mental model

5. DISCUSSION

Natural Trust

- The experiment showed a relationship between the inclusion of the summary and **natural trust.**
- Results could be attributed to transparency and explainability.

Overall Satisfaction

- The experiment revealed a correlation between the inclusion of the summary and **overall satisfaction**.
- Results could be attributed to transparency, and "gamification".

Performance

- Increased **artificial trust** supports the notion of the feedback loop.

Limitations & Future Work

- Increase the sample size.
- Consider different contexts.
- Longitudinal studies.

REFERENCES

[1] J E Hans Korteling et al. "Human- versus Artificial Intelligence". en. In: Front. Artif. Intell. 4 (Mar. 2021), p. 622364.
[2] Mayer, R. C., Davis, J. H., & Schoorman, F. D. (2020). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335
[3] Carolina Centeio Jorge, Emma, Verhagen, R., Mehrotra, S., Jonker, C. M., & Tielman, M. L. (2024). Appropriate context-dependent artificial trust in human-machine teamwork. Elsevier EBooks, 41–60. https://doi.org/10.1016/b978-0-443-15988-6.00007-8
[4] Andrews, R. W., Lilly, J. M., Srivastava, D., & Feigh, K. M. (2022). The role of shared mental models in human-AI teams: a theoretical review. Theoretical Issues in Ergonomics Science, 24(2), 1–47. https://doi.org/10.1080/1463922x.2022.2061080
[5] Matthew B. Luebbers*, Aaquib Tabrez*, Kyler Ruvane*, and Bradley Hayes. (2023). Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming. In Proceedings of Robotics: Science and Systems (RSS 2023).