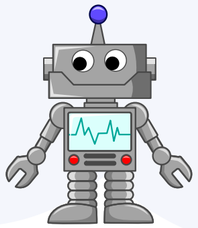


Communicating trust-based beliefs and decisions in human AI-teams

Author: Tamer Sahin (t.sahin@student.tudelft.nl)
Supervisor: Carolina Jorge
Responsible Professor: Myrthe Tielman

1 Background

Human-AI teams (HATs) [1]
Teams that consist of a human and an artificial agent working jointly or individually on tasks to achieve a common goal.

Trust: [2,3]
The trust between the trustor and trustee is dynamic and can change based on the actions and behavior of the trustee.

- **Mutual trust:** humans and artificial agents rely on each other to do their tasks well.
- **Artificial trust:** the trust from the agent towards the human.
- **Natural trust:** the trust from the human towards the agent.

The artificial agent measures the artificial trust through competence and willingness:

- **Competence:** to help decide whether it should believe that the human could be useful for the action.
- **Willingness:** to help decide whether it should believe that the human would perform the action as expected.

Figure 1: Screenshot of the environment

2 Research Question

How does a **real-time textual explanation** of the **mental model** of the agent's trust in the human teammate affect the **human teammate's trust** in the agent and overall **satisfaction**?

3 Trust Model

- Mental model tracks the **competence** and **willingness** of the 3 tasks: search/victims/obstacles.
- **Preference Integration** to measure the willingness for certain tasks:
  - Flooded vs non-flooded
  - Difficult victims vs non-difficult victims
  - Distance: far vs close

4 The Experiment

- The human collaborates with the agent (RescueBot) in a search and rescue mission. (Figure 1)
- The experiment compared a standard baseline model against a real-time textual explanation model to evaluate the effectiveness of the new approach.
- Real-time textual explanations provided insights into why the RescueBot behaved in certain ways and also indicated whether the artificial trust increased or decreased. (Figure 2)

Subjective measures are measured through questionnaires:

- Trust and satisfaction

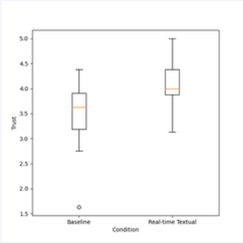
Objective measures are logged automatically:

- Artificial trust (competence and willingness)
- Score
- Completeness

5 Results

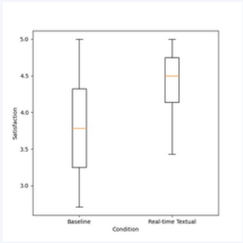
Natural Trust

- The Shapiro-Wilk test indicates that both datasets are normally distributed
- Both datasets satisfy the assumption of homogeneity of variance as measured by Levene's test.
- A **t-test** is conducted where a significant difference is found ($p \approx 0.002$).



Satisfaction

- The Shapiro-Wilk test indicates that both datasets are normally distributed
- Both datasets satisfy the assumption of homogeneity of variance as measured by Levene's test.
- A **t-test** is conducted where a significant difference is found ($p \approx 0.002$).



Artificial Trust

- A standard **t-test** is conducted.
- No statiscally significant differences ($p = 0.188$).

References
[1] C. Flathmann, B. G. Schelble, R. Zhang, and N. J. Mc- Neese, "Modeling and guiding the creation of ethical human-ai teams," pp. 469–479, 2021
[2] A. C. Costa, R. A. Roe, and T. Taillieu, "Trust within teams: The relation with performance effectiveness," European journal of work and organizational psychology, vol. 10, no. 3, pp. 225–244, 2001.
[3] H. Azevedo-Sa, X. J. Yang, L. P. Robert, and D. M.Tilbury, "A unified bi-directional model for natural and artificial trust in human–robot collaboration," IEEE robotics and automation letters, vol. 6, no. 3, pp. 5913–5920, 2021

6 Discussion

Natural Trust

- Results show that real-time textual explanations enhance the natural trust of the human.
- Potential factors for the increase could be transparency and predictability.

Satisfaction

- Results show that real-time textual explanations enhance the satisfaction of the human.
- Potential factors for the increase could be emotional engagement and good player intentions.

Artifical Trust

- No definitive conclusion can be drawn as the results are not statistically significant.

Limitations and Future Work

- All of the participants resided in Europe.
- The experiment was conducted with 40 participants.
- Compare another real-time textual explanation model with the baseline.

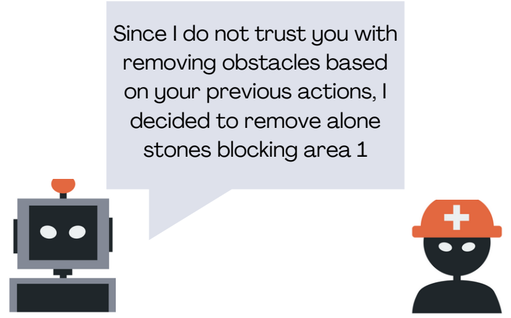


Figure 2: RescueBot communicating with real-time textual explanations