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

AUTHORS

Author: Elena Uleia (E.Uleia@student.tudelft.nl)
Supervisor: Carolina Centeio Jorge
Responsible Professor: Myrthe Tielman

TU Delft

1. Introduction and Background

Motivation: artificial intelligence evolved from a tool to a human counterpart in teams **[1]**

- **artificial trust:** from AI toward human
- **natural trust:** from humans toward AI **[2]**
- **mental model:** an internal representation of external reality
- **competence:** can a human provide the expected result?
- **willingness:** does the human want to achieve the expected result?
- **preference modeling:** what tasks do humans prefer?

Communication

- **real-time** (timing)**:** proactiveness enhances collaboration **[3]**
- **visual** (content type): enhances collaboration most compared to other methods (verbal and audio) **[4]**

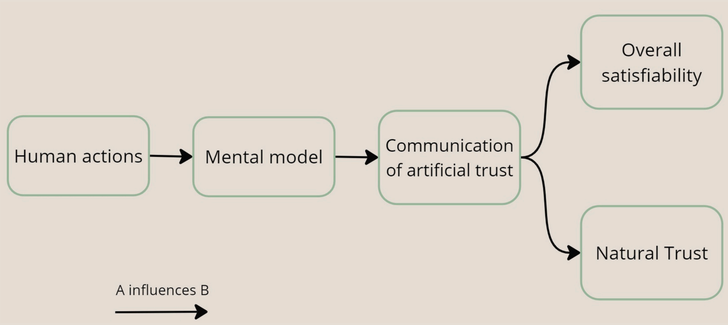

Figure 1: The flow of communication impact on trust and satisfiability

2. Research Question

How does a **real-time visual** (RTV) communication of the mental model of the agent's trust affect the human teammate's **trust** in the agent and overall **satisfaction**?

3. Trust Mental Model

Trust matrix (3 x 2)

- (C, W) per task
- task types: search, obstacles, victims

Value updates

- **p** preference modeling factor (distance, speed)

$$\begin{cases} W_{task}(t) = W_{task}(t-1) \pm I + p \\ C_{task}(t) = C_{task}(t-1) \pm I \end{cases}$$

Figure 2: Picture of the environment

Thresholds

u ~ U(0,1) confidence in own decisions

$u < confidence$ (1) $\quad W_{task} \geq \frac{1-p}{2}$ (2) $\quad C_{task} \geq 0$ (3)

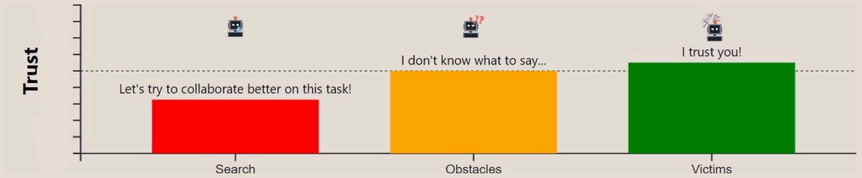

Figure 3: Representation of the visual communication of the mental model

4. Methodology

- In-between **user experiment**
- Experiment groups: Baseline(n=22) and RTV Communication(n=22)
- Measurements trust: **questionnaire** and behavioral measures (compliance, collaboration frequency, **game duration**, number of human messages, artificial trust)
- Measurements overall satisfaction: **questionnaire**

5. Results

- Mann-Whitney U test to compare Baseline and RTV

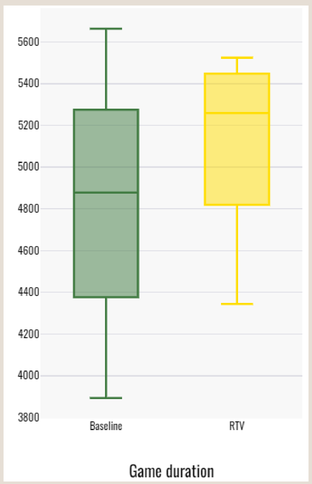

*Figure 4: Boxplot comparing **game duration** between Baseline and RTV Communication conditions*

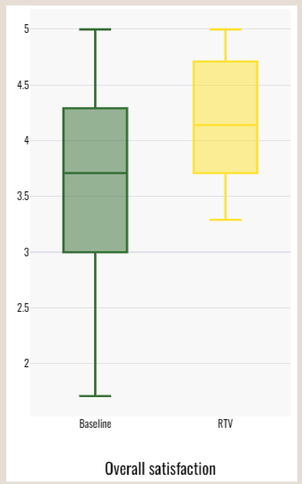

*Figure 5: Boxplot comparing **overall satisfaction** between Baseline and RTV Communication conditions*

Limitations

- different devices with different performance
- participants' reduced diversity
- environmental setup

Future Work

- fine-tune the hyperparameters of the trust model
- compare the effect of communication on trust between computer scientists and non-computer scientists

6. Conclusion and Discussion

- overall satisfaction results align with previous research **[5]**

1. Trust

- insignificant results due to participants' AI knowledge
- baseline participants were faster due to laptop increased performance and lower cognitive load

2. Overall satisfaction

- RTV communication **improves** overall satisfaction

References

[1] L. Larson and L. A. DeChurch, 'Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams', The leadership quarterly, vol. 31, no. 1, p. 101377, 2020.
[2] C. C. Jorge, C. M. Jonker, and M. L. Tielman, 'Artificial trust for decision-making in human-AI teamwork: Steps and challenges', in Proceedings of the HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI), 2023.
[3] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. Investi- gating ai teammate communication strategies and their impact in human-ai teams for effective teamwork. Pro-ceedings of the ACM on Human-Computer Interaction, 7(CSCW2):1–31, 2023
[4] Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-ai collaboration. Plos one, 15(2):e0229132, 2020.
[5] Vijai N Giri and B Pavan Kumar. Assessing the impact of organizational communication on job satisfaction and job performance. Psychological Studies, 55:137 143, 2010.