Communicating trust-based beliefs and decisions in human-AI teams using real-time visual explanations

Supervisor: Carolina Centeio Jorge

Author: Elena Dumitrescu (E.I.Dumitrescu-1@student.tudelft.nl)

Responsible professor: Myrthe Tielman

1. INTRODUCTION

Human-AI teams need **mutual trust** to collaborate effectively:

- **natural trust** = human's trust in AI agent
- **artificial trust** = AI agent's trust in human

Artificial trust: perceive human characteristics and assess whether they are a cue for trustworthiness ⇒ **mental model** of the human

Knowledge gap: Little empirical research and implementations of artificial trust models

Communication: necessary to establish natural trust. The type (e.g. visual/textual) impacts team trust and performance [1].

Knowledge gap:

- **No studies on the advantages communicating artificial trust**
- **Little focus on the different types of communication human-AI teams can incorporate [1]**

2. RESEARCH QUESTION

How do **real-time visual explanations** of the mental model of the **AI agent's trust** in its human teammate affect the human's **trust** in the AI agent and overall **satisfaction**?

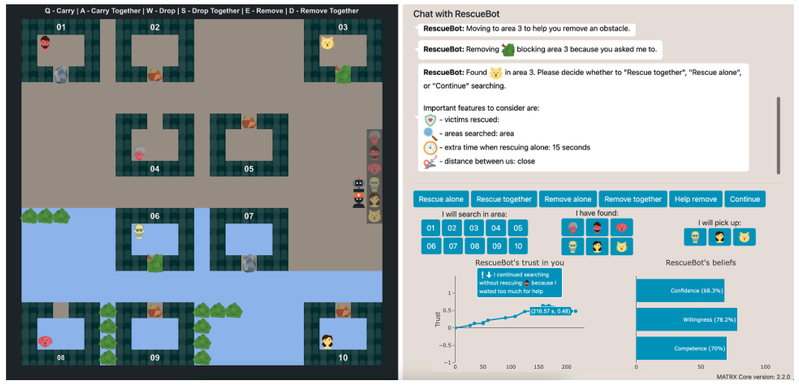

Figure 1. Environment. Top left shows the initial map configuration. Top right shows the chat area. Bottom presents a zoomed image of the trust graphs (for TE group only).

3. ARTIFICIAL TRUST PROCESS

Evaluation: AI adjusts (Δ) its beliefs (B) about human competence and willingness (ϕ) on task $t \in D$, based on behaviour cues (n)

$$\begin{cases} B_n(\phi(H,D)) = B_{n-1}(\phi(H,D)) + \Delta(t) + P(t) \\ B_0(\phi(H,D)) = 0 \end{cases} \quad (1)$$

Decision: after trust evaluation, decide (τ) whether to trust the human ⇒ possibly **adapt behaviour** based on decision

$$\tau_n(t) = B_n(competent(H,D)) \geq T_c \wedge B_n(willing(H,D)) \geq T_w \quad (2)$$

Context: integrated in both evaluation and decision, e.g. how preferable is a task (P)

4. METHOD

Between-subject **experiment** with 46 participants, comparing trust explanations group (TE) against baseline. Human and AI collaborate to save 6 victims in a **search and rescue** task (Figure 1)

Communication: on every trust/behaviour update

- **Time-based plot** - aggregated trust value over time, explanations for each data point.
- **Beliefs bar chart** - AI agent's beliefs

Preference integration: heuristic-based

- **flooded areas** (longer to navigate)
- **special victims** (longer to rescue)
- **distance** (human prefers closer tasks)

Subjective measures: self-reported trust and satisfaction, measured with Likert scale questionnaires

Objective measures: communication rate, level of interaction with robot, mouse movements, focus on trust plots, compliance

5. RESULTS - STATISTICAL TESTS

H1/H2 *Incorporating real-time visual explanations of the AI agent's trust in its human teammate increases* ***natural trust/overall satisfaction****.*

Pearson's correlations between subjective and objective measures (Table 1)

Comparison tests between dependent variables across the two conditions (Table 2). Parametric assumptions verified beforehand.

Mouse movements heatmap aggregated for all participants in the TE group (Figure 2)

$\alpha = 0.05$

Table 1. Pearson's correlations between self-reported measures and objective metrics

	Communication Rate	Level of Interaction	Focus	Compliance
SR Satisfaction	0.34*	-0.287	0.095	-0.423*
SR Trust	0.247	-0.113	0.182	-0.088

* Statistically significant at $p < 0.05$ level (green).

Table 2. Comparison test results for assessing differences across the two conditions

Metric	Statistical Test	P-value	Condition	Mean (μ)	SD (σ)
SR Trust	Independent Samples Welch's T-test	< 0.001*	TE	4.261	0.31◇
			Baseline	3.511	0.624◇
SR Satisfaction	Independent Samples Welch's T-test	0.002*	TE	4.344	0.41◇
			Baseline	3.688	0.873◇
Communication rate	Mann-Whitney U test	0.011*	TE	0.049	0.011
			Baseline†	0.042	0.014
Compliance	Mann-Whitney U test	0.216	TE†	2.864	1.66
			Baseline	3.09	1.311

* Statistically significant at $p < 0.05$ level (green).
† Non-normality (orange).
◇ Heteroscedasticity (yellow).

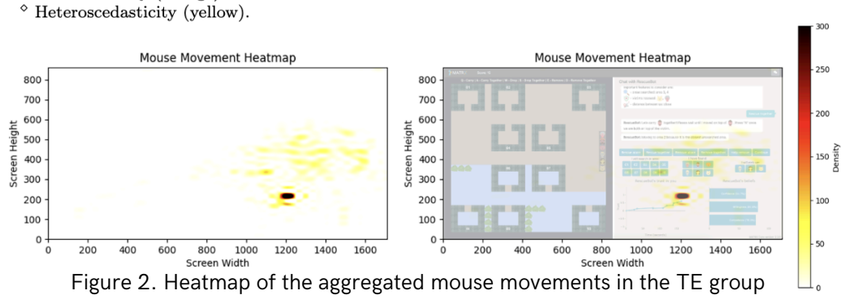

Figure 2. Heatmap of the aggregated mouse movements in the TE group

6. DISCUSSION

Self-reported trust: including real-time visual explanations increases trust, *supporting H1*

Self-reported satisfaction: including real-time visual explanations increases satisfaction, *supporting H2*

Communication rate: positively correlated with satisfaction and higher for the TE group, *supporting H2*

Compliance: negatively correlated with satisfaction ⇒ measure of perceived task difficulty, as it also shows dependence on AI

Mouse movements heatmap shows general interest in hovering functionality, however participants' opinions also reveal potential information overload

Limitations: hardware inconsistencies, homogeneity of participants

Future work:

- compare different trust communication types, including hybrid
- explore more metrics for trust and satisfaction
- generally focus on empirical research for artificial trust

[1] Rui Zhang et al. "Investigating AI Teammate Communication Strategies and Their Impact in Human-AI Teams for Effective Teamwork". In: Proceedings of the ACM on Human-Computer Interaction 7 (2023), pp. 1–31